Data Mining and Knowledge Discovery (2024) 38:2997–3042
https://doi.org/10.1007/s10618-023-00993-x

Marginal effects for non-linear prediction functions

Christian A. Scholbeck[1] · Giuseppe Casalicchio[1] · Christoph Molnar[2] · Bernd Bischl[1] · Christian Heumann[2]

Received: 31 March 2021 / Accepted: 23 November 2023 / Published online: 27 February 2024

Abstract

Beta coefficients for linear regression models represent the ideal form of an interpretable feature effect. However, for non-linear models such as generalized linear models, the estimated coefficients cannot be interpreted as a direct feature effect on the predicted outcome. Hence, marginal effects are typically used as approximations for feature effects, either as derivatives of the prediction function or forward differences in prediction due to changes in feature values. While marginal effects are commonly used in many scientific fields, they have not yet been adopted as a general model-agnostic interpretation method for machine learning models. This may stem from the ambiguity surrounding marginal effects and their inability to deal with the non-linearities found in black box models. We introduce a unified definition of forward marginal effects (FMEs) that includes univariate and multivariate, as well as continuous, categorical, and mixed-type features. To account for the non-linearity of prediction functions, we introduce a non-linearity measure for FMEs. Furthermore, we argue against summarizing feature effects of a non-linear prediction function in a single metric such as the average marginal effect. Instead, we propose to average homogeneous FMEs within population subgroups, which serve as conditional feature effect estimates.

Keywords Forward difference · Forward marginal effects · FME · NLM · cAME · Non-linearity measure · Conditional average marginal effects · Interpretable machine learning · Explainable AI · IML · XAI · Interpretations · Model-agnostic

Responsible editor: Martin Atzmüller, Johannes Fürnkranz, Tomas Kliegr, and Ute Schmid.

✉ Christian A. Scholbeck
christian.scholbeck@stat.uni-muenchen.de

Giuseppe Casalicchio
giuseppe.casalicchio@stat.uni-muenchen.de

[1] Munich Center for Machine Learning (MCML), Ludwig-Maximilians-Universität in Munich, Munich, Germany

[2] Ludwig-Maximilians-Universität in Munich, Munich, Germany

1 Introduction

The lack of interpretability of most machine learning (ML) models has been considered one of their major drawbacks (Breiman 2001b). As a consequence, researchers have developed a variety of model-agnostic techniques to explain the behavior of ML models. These techniques are commonly referred to by the umbrella terms of interpretable machine learning (IML) or explainable artificial intelligence. Model explanations take different forms, e.g., feature attributions (FAs) such as a value indicating a feature's importance to the model or a curve indicating its effects on the prediction, model internals such as beta coefficients for linear regression models, data points such as counterfactual explanations (Wachter et al. 2018), or surrogate models (i.e., interpretable approximations to the original model) (Molnar 2022). In the context of our paper, we categorize an FA as an effect or importance:

- **Feature effect:** We define a feature effect as the direction and magnitude of a change in predicted outcome due to a change in feature values (Casalicchio et al. 2019; Scholbeck et al. 2020).
- **Feature importance:** Importance is an indication of a feature's relevance to the model. Effect and importance are related, as a feature with a large effect on the prediction can also be considered important. However, a feature's relevance can be measured in multiple ways; for instance, the permutation feature importance (Fisher et al. 2019) shuffles feature values and evaluates changes in model performance, while the functional analysis of variance (Saltelli et al. 2008; Hooker 2004b, 2007) evaluates contributions of terms within a high-dimensional model representation to the model output variance.

In this paper, we focus on feature effects, which are relevant for many applications. We distinguish between local explanations on the observational level and global ones for the entire feature space. For example, in medical research, we might want to assess the increase in risk of contracting a disease due to a change in a patient's health characteristics such as age or body weight. Consider the interpretation of a linear regression model (LM) without interaction terms where β_j denotes the coefficient of the j-th feature. Increasing a feature value x_j by one unit causes a change in predicted outcome of β_j. LMs are therefore often interpreted by merely inspecting the estimated coefficients. When the terms are non-linear, interactions are present, or when the expected target is transformed such as in generalized linear models (GLMs), interpretations are both inconvenient and unintuitive. For instance, in logistic regression, the expectation of the target variable is logit-transformed, and the predictor term cannot be interpreted as a direct feature effect on the predicted risk. It follows that even linear terms have a non-linear effect on the predicted target that varies across the feature space and makes interpretations through the model parameters difficult to impossible. A more convenient and intuitive interpretation corresponds to the derivative of the prediction function w.r.t. the feature or inspecting the change in prediction due to an intervention in the data. These two approaches are commonly referred to as marginal effects (MEs) in statistical literature (Bartus 2005). MEs are often aggregated to an average marginal effect (AME), which represents an estimate of the expected ME. Furthermore, marginal effects at means (MEM) and marginal effects at representative values (MER) correspond to MEs where all features are set to the sample mean or where some feature values are set to manually chosen values (Williams 2012). These can be used to answer common research questions, e.g., what the average effect of age or body weight is on the risk of contracting the disease (AME), what the effect is for a patient with average age and body weight (MEM), and what the effect is for a patient with pre-specified age and body weight values (MER). An increasing amount of scientific disciplines now rely on the predictive power of black box ML models instead of using intrinsically interpretable models such as GLMs, e.g., econometrics (Athey 2017) or psychology (Stachl et al. 2017). This creates an incentive to review and refine the theory of MEs for the application to non-linear models.

For one, there is much confusion regarding the definition of MEs, evidenced by two variants for continuous features (based on either derivatives or forward differences) and furthermore by categorical MEs (which are computed as finite differences resulting from switching categories in various ways). In their current form, MEs are not an ideal tool to interpret many statistical models such as GLMs, and their shortcomings are exacerbated when applied to black box models such as the ones created by many ML algorithms. For non-linear prediction functions, MEs based on derivatives provide misleading feature effect interpretations: Given the tangent to the prediction function at a point x, we evaluate the tangent's rise at a point $x + h$. A unit increase for h is typically used as an interpretable standard measure. For non-linear prediction functions however, this change in feature values results in a different prediction than implied by the derivative ME, thereby rendering this interpretation misleading. The alternative and often overlooked definition based on forward differences is much better suited for effect interpretations but also suffers from a loss in information about the shape of the prediction function (see Sect. 3). For linear models, the ME is identical across the entire feature space. For non-linear models, one typically estimates the global feature effect by computing the AME (Bartus 2005; Onukwugha et al. 2015). However, a global average does not accurately represent the nuances of a non-linear predictive model. A more informative summary of the prediction function corresponds to the conditional feature effect on a feature subspace, e.g., patients with an entire range of health characteristics might be associated with homogeneous feature effects. Instead of global interpretations on the entire feature space, one should instead aim for semi-aggregated (regional or semi-global) interpretations. More specifically, one should work towards computing multiple, regional conditional AMEs (cAMEs) instead of a single, global AME.

Contributions: This paper introduces forward marginal effects (FMEs) as a model-agnostic interpretation method for arbitrary prediction functions[1]. We first provide a unified definition of FMEs for both univariate and multivariate, as well as continuous, categorical, and mixed-type features. Then, we define a non-linearity measure (NLM) for FMEs based on the similarity between the prediction function and the intersecting linear secant. Furthermore, for a more nuanced interpretation, we introduce conditional AMEs (cAMEs) for population subgroups as a regional (semi-global) feature effect measure that more accurately describes feature effects across the feature space. We propose one option to find subgroups for cAMEs by recursively partitioning the feature space with a regression tree on FMEs. Furthermore, we provide proofs on additive recovery for the univariate and multivariate FME and a proof on the relation between the individual conditional expectation (ICE) / partial dependence (PD) and the FME / forward AME.

Structure of the paper: In Sect. 2, we introduce our notation. In Sect. 3, we make sense of the ambiguous usage of MEs. In Sect. 4, we introduce a unified definition of FMEs, the NLM, and the cAME. Section 5 provides an overview on related work, demonstrates the relation between FMEs and the ICE / PD, and compares FMEs to the competing approach LIME. In Sect. 6, we run multiple simulations showcasing FMEs and the NLM. In Sect. 7, we present a structured application workflow and an applied example on real data. The Appendix contains background information on additive decompositions of prediction functions, on model extrapolations, on MEs for tree-based functions, as well as the above-mentioned mathematical proofs.

2 Notation

We consider a p-dimensional feature space $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_p$ and a target space $\mathscr{Y}$. The random variables on the feature space are denoted by $\boldsymbol{X} = (X_1, \ldots, X_p)$.[2] The random variable on the target space is denoted by Y. A generic subspace of all features is denoted by $\mathscr{X}_{[\,]} \subseteq \mathscr{X}$. Correspondingly, $\boldsymbol{X}$ with a restricted sample space is denoted by $\boldsymbol{X}_{[\,]}$. A realization of $\boldsymbol{X}$ and Y is denoted by $\boldsymbol{x} = (x_1, \ldots, x_p)$ and y. The probability distribution $\mathscr{P}$ is defined on the sample space $\mathscr{X} \times \mathscr{Y}$. A learning algorithm trains a predictive model $\widehat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ on data drawn from $\mathscr{P}$, where $\widehat{f}(\boldsymbol{x})$ denotes the model prediction based on the p-dimensional feature vector $\boldsymbol{x}$. To simplify our notation, we only consider one-dimensional predictions. However, the results on MEs can be generalized to multi-dimensional predictions, e.g., for multi-class classification. We denote the value of the j-th feature in $\boldsymbol{x}$ by x_j. A set of features is denoted by $S \subseteq \{1, \ldots, p\}$. The values of the feature set are denoted by $\boldsymbol{x}_S$.[3] All complementary features are indexed by $-j$ or $-S$, so that $\boldsymbol{x}_{-j} = \boldsymbol{x}_{\{1, \ldots, p\} \setminus \{j\}}$, or $\boldsymbol{x}_{-S} = \boldsymbol{x}_{\{1, \ldots, p\} \setminus S}$. An instance $\boldsymbol{x}$ can be partitioned so that $\boldsymbol{x} = (x_j, \boldsymbol{x}_{-j})$, or $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{-S})$. With slight abuse of notation, we may denote the vector $\boldsymbol{x}_S$ by $(x_1, \ldots, x_s)$ regardless

[1] During the peer review process, we began to implement the theory presented in this manuscript in the R package `fmeffects` (Löwe et al. 2023)

[2] Vectors are denoted in bold letters.

[3] As $\boldsymbol{x}_S$ is the generalization of x_j to vectors, we denote it in bold letters. However, it can in fact be a scalar. The same holds for $\boldsymbol{x}_{-S}$ and $\boldsymbol{x}_{-j}$.

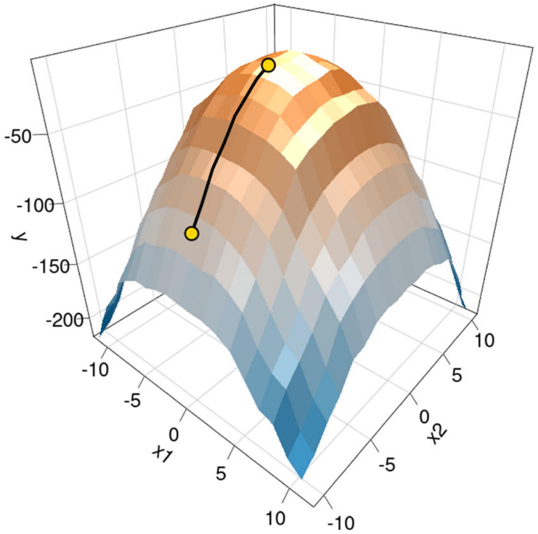


Fig. 1 The surface represents an exemplary prediction function dependent on two features. The FD can be considered a movement on the prediction function. We travel from point (0, −9) to point (0, −2) (Color figure online)

of the elements of S, or the vector $(x_j, \boldsymbol{x}_{-j})$ by $(x_1, \ldots, x_j, \ldots, x_p)$ although $j \in \{1, \ldots, p\}$. The i-th observed feature vector is denoted by $\boldsymbol{x}^{(i)}$ and corresponds to the target value $y^{(i)}$. We evaluate the prediction function with a set of training or test data $\mathscr{D} = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$.

A finite difference (FD) of the prediction $\widehat{f}(\boldsymbol{x})$ w.r.t. x_j is defined as:

$$\mathrm{FD}_{j,\boldsymbol{x},a,b} = \widehat{f}(x_1, \ldots, x_j + a, \ldots, x_p) - \widehat{f}(x_1, \ldots, x_j + b, \ldots, x_p)$$

The FD can be considered a movement on the prediction function (see Fig. 1). There are three common variants of FDs: forward ($a = h$, $b = 0$), backward ($a = 0$, $b = -h$), and central differences ($a = h, b = -h$). In the following, we only consider forward differences with $b = 0$ where the FD is denoted without b. Dividing the FD by $(a - b)$ corresponds to the difference quotient:

$$\frac{\mathrm{FD}_{j,\boldsymbol{x},a,b}}{a - b} = \frac{\widehat{f}(x_1, \ldots, x_j + a, \ldots, x_p) - \widehat{f}(x_1, \ldots, x_j + b, \ldots, x_p)}{a - b}$$

The derivative is defined as the limit of the forward difference quotient when $a = h$ approaches zero:

$$\frac{\partial \widehat{f}(\boldsymbol{X})}{\partial X_j}\bigg|_{\boldsymbol{X}=\boldsymbol{x}} = \lim_{h \to 0} \frac{\widehat{f}(x_1, \ldots, x_j + h, \ldots, x_p) - \widehat{f}(\boldsymbol{x})}{h}$$

We can numerically approximate the derivative with small values of h. For instance, we can use forward, backward, or symmetric FD quotients, which have varying error characteristics. As an example, consider a central FD quotient which is often used for derivative-based MEs (Leeper 2018):

$$\frac{\partial \widehat{f}(\boldsymbol{X})}{\partial X_j}\bigg|_{\boldsymbol{X}=\boldsymbol{x}} \approx \frac{\widehat{f}(x_1, \ldots, x_j + h, \ldots, x_p) - \widehat{f}(x_1, \ldots, x_j - h, \ldots, x_p)}{2h} \quad , h > 0$$

3 Making sense of marginal effects

There is much ambiguity and confusion surrounding MEs. They are either defined in terms of derivatives or forward differences, and there is further confusion regarding the definition of categorical MEs.

3.1 Marginal effects for categorical features

MEs for categorical features are often computed as the change in prediction when the feature value changes from a reference category to another category (Williams 2012). In other words, for each observation, the observed categorical feature value is set to the reference category, and we record the change in prediction when changing it to every other category. Given k categories, this results in $k - 1$ MEs for each observation. Consider a categorical feature indexed by j with categories $C = \{c_1, \ldots, c_k\}$. We select a reference category $c_r \in C$. The categorical ME for an observation $\boldsymbol{x}$ and a single category $c_l \in C \setminus \{c_r\}$ corresponds to:

$$\text{ME}_{j,\boldsymbol{x},c_r,c_l} = \widehat{f}(c_l, \boldsymbol{x}_{-j}) - \widehat{f}(c_r, \boldsymbol{x}_{-j})$$

3.2 Marginal effects for continuous features

3.2.1 Definition as derivative

The most commonly used definition of MEs for continuous features corresponds to the derivative of the prediction function w.r.t. a feature. We will refer to this definition as the derivative ME (DME). In case of a linear prediction function, the interpretation of DMEs is simple: if the feature value increases by one unit, the prediction will increase by the DME estimate. Note that even the prediction function of a linear regression model can be non-linear if exponents of order ≥ 2 are included in the feature term. Similarly, in GLMs, the linear predictor is transformed (e.g., log-transformed in Poisson regression or logit-transformed in logistic regression).

3.2.2 Definition as forward difference

A distinct and often overlooked definition of MEs corresponds to the change in prediction with adjusted feature values, also referred to as discrete change (Mize et al. 2019) or difference in adjusted predictions (APs) (Williams 2012). This definition of MEs is based on a forward difference instead of a symmetric difference and does not require dividing the FD by the interval width. For this reason—and to establish a unified definition of MEs—we refer to this variant as the forward ME (FME):

$$\begin{aligned} \text{FME}_{\boldsymbol{x},\boldsymbol{h}_S} &= \widehat{f}(x_1 + h_1, \ \ldots \ , x_s + h_s, \boldsymbol{x}_{-S}) - \widehat{f}(x_1, \ \ldots \ , x_s, \boldsymbol{x}_{-S}) \\ &= \widehat{f}(\boldsymbol{x}_S + \boldsymbol{h}_S, \boldsymbol{x}_{-S}) - \widehat{f}(\boldsymbol{x}) \end{aligned} \tag{1}$$

A univariate FME for $h = 1$ is illustrated in Fig. 2. It corresponds to the change in prediction along the secant (orange, dotdashed) through the point of interest (prediction at x = 0.5) and the prediction at the feature value we receive after the feature change (x = 1.5).

Note that FMEs—as any other model-agnostic method—may result in model extrapolations if based on predictions in areas where the model was not trained with a sufficient amount of data. In Appendix A.2 , we discuss model extrapolations and how they relate to the computation of FMEs.

A technique that is subject to the additive recovery property only *recovers* terms of the prediction function that depend on the feature(s) of interest $\boldsymbol{x}_S$ or consist of interactions between the feature(s) of interest and other features, i.e., the method recovers no terms that exclusively depend on the remaining features $\boldsymbol{x}_{-S}$ (Apley and Zhu 2020). In Appendix B. 1, we derive the additive recovery property for FMEs.

3.2.3 Forward difference versus derivative

Note that we refer to using MEs to obtain **feature effect interpretations** (see Sect. 1), meaning changes in predicted outcome due to changes in feature values (locally and globally). In case of non-linear prediction functions, using DMEs for effect interpretations can lead to substantial misinterpretations (see Fig. 2). The slope of the tangent (green, dashed) at the point of interest (prediction at x = 0.5) corresponds to the DME. The default way to obtain a feature effect using the DME is to evaluate the tangent at the feature value we receive **after** changing feature values (in this case, we make a unit change, resulting in $x = 1.5$). This leads to substantial misinterpretations for non-linear prediction functions. In this case, there is an error (purple) almost as large as the actual change in prediction (the FME, blue). Although the computation of the DME does not require a step size, its interpretation does and is therefore error-prone. In contrast, the FME always indicates an exact change in prediction for any prediction function and is therefore much more interpretable. Only for linear prediction functions, the interpretation of both variants is equivalent.

There is a further advantage of FMEs over DMEs: derivatives are not suited to interpret piecewise constant prediction functions such as the ones created by tree-based algorithms. We discuss this point in more detail in Appendix A.3.

3.3 Variants and aggregations of marginal effects

There are three established variants or aggregations of MEs: The AME, MEM, and MER (Williams 2012), which can be computed for both DMEs and FMEs. In the following, we will use the notation of FMEs. Although we technically refer to the FAME, FMEM, and FMER, we omit the “forward” prefix in this case for reasons of simplicity:

(i) **Average marginal effect (AME):** The AME represents an estimate of the expected FME w.r.t. the distribution of $\boldsymbol{X}$. We estimate it via Monte-Carlo integration, i.e., we average the FMEs that were computed for each (randomly sampled) observation:

$$\mathbb{E}_{\boldsymbol{X}}\left[\text{FME}_{\boldsymbol{X},\boldsymbol{h}_S}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\widehat{f}(\boldsymbol{X}_S + \boldsymbol{h}_S, \boldsymbol{X}_{-S}) - \widehat{f}(\boldsymbol{X})\right]$$

$$\text{AME}_{\mathscr{D},\boldsymbol{h}_S} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{f}\left(\boldsymbol{x}_S{}^{(i)} + \boldsymbol{h}_S, \boldsymbol{x}_{-S}^{(i)}\right) - \widehat{f}\left(\boldsymbol{x}^{(i)}\right)\right]$$

(ii) **Marginal effect at means (MEM):** The MEM can be considered the reverse of the AME, i.e., it is the FME evaluated at the expectation of $\boldsymbol{X}$. We estimate the MEM by replacing all feature values with their sample distribution means:

$$\text{FME}_{\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}],\boldsymbol{h}_S} = \widehat{f}\left(\mathbb{E}_{\boldsymbol{X}_S}[\boldsymbol{X}_S] + \boldsymbol{h}_S, \mathbb{E}_{\boldsymbol{X}_{-S}}\left[\boldsymbol{X}_{-S}\right]\right) - \widehat{f}\left(\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}]\right)$$

$$\text{MEM}_{\mathscr{D},\boldsymbol{h}_S} = \widehat{f}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_S^{(i)}\right) + \boldsymbol{h}_S, \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{-S}^{(i)}\right) - \widehat{f}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)}\right)$$

Note that averaging values is only sensible for continuous features. Williams (2012) defines a categorical MEM where all remaining features are set to their sample means (conditional on being continuous) and the feature of interest changes from a reference category to every other category.

(iii) **Marginal effect at representative values (MER):** Furthermore, we can replace specific feature values for all observations with manually specified values $\boldsymbol{x}^*$. It follows that the MEM is a special case of the MER where the specified values correspond to the sample means. MERs can be considered conditional FMEs, i.e., we compute FMEs while conditioning on certain feature values. The MER for a single observation with modified feature values $\boldsymbol{x}^*$ corresponds to:

$$\text{MER}_{\boldsymbol{x}^*,\boldsymbol{h}_S} = \widehat{f}\left(\boldsymbol{x}_S^* + \boldsymbol{h}_S, \boldsymbol{x}_{-S}^*\right) - \widehat{f}\left(\boldsymbol{x}^*\right)$$

The AME, MEM, and MER are mainly targeted at continuous features. In Sect. 4, we discuss computations for unified FMEs.

4 Model-agnostic forward marginal effects for arbitrary prediction functions

4.1 Unified definition of forward marginal effects

Note that both categorical MEs and FMEs are based on forward differences. We propose a unified definition of FMEs for continuous, categorical, and mixed-type features in S. Recall that the definition of FMEs for continuous features is given by Eq. (1):

$$\text{FME}_{\boldsymbol{x},\boldsymbol{h}_S} = \widehat{f}(\boldsymbol{x}_S + \boldsymbol{h}_S, \boldsymbol{x}_{-S}) - \widehat{f}(\boldsymbol{x}) \qquad \text{for continuous features } \boldsymbol{x}_S$$

We suggest an observation-specific categorical FME, where we first select a single category c_j and predict once with the observed value x_j and once where x_j has been replaced by c_j:

$$\text{FME}_{\boldsymbol{x},c_j} = \widehat{f}(c_j, \boldsymbol{x}_{-j}) - \widehat{f}(\boldsymbol{x}) \qquad \text{for categorical } x_j$$

This definition of categorical FMEs is congruent with the definition of FMEs for continuous features, as we receive a single FME for a single observation with the observed feature value as the reference point. In other words, the reference category c_j for a categorical FME is conceptually identical to the step size h_j for a continuous FME. This implies that for observations where $x_j = c_j$, the categorical FME is zero. Continuous and categorical FMEs can be combined for mixed-data FMEs. Consider a set $S = \{j, l\}$ and the vector $\boldsymbol{h}_S = (h_j, c_l)$ with step size h_j for the j-th feature (which is continuous) and a category c_l for the l-th feature (which is categorical). A mixed-type FME is given by:

$$\text{FME}_{\boldsymbol{x},\boldsymbol{h}_S} = \widehat{f}(x_j + h_j, c_l, \boldsymbol{x}_{-S}) - \widehat{f}(\boldsymbol{x}) \qquad \text{for continuous } x_j \text{ and categorical } x_l$$

We therefore remove any ambiguity from MEs through a unified definition and terminology based on forward differences for all feature types.

Categorical FMEs and the computation of MEMs and MERs: Categorical FMEs are also suited for computing a categorical AME. Note that we generally have less than n categorical FMEs different from zero, depending on the observed marginal distribution of $\{x_j^{(i)}\}_{i=1}^n$, which may affect the variance of the mean. Although the computation of MERs for categorical FMEs is possible, the MER obfuscates their interpretation by destroying the empirical distribution.

4.2 Non-linearity measure for continuous features

Although an FME represents the exact change in prediction and always accurately describes the movement on the prediction function, we lose information about the function's shape along the forward difference. It follows that when interpreting FMEs, we are at risk of misjudging the shape of the prediction function as a piecewise linear function. However, prediction functions created by ML algorithms are not only non-linear but also differ considerably in shape across the feature space. We suggest to augment the change in prediction with an NLM that quantifies the deviation between the prediction function and a linear reference function. First, the FME tells us the change in prediction for pre-specified changes in feature values. Then, the NLM tells us how accurately a linear effect resembles the change in prediction. The NLM thus represents a measure of confidence whether interpolations regarding the FME along the step are possible. For instance, assume the associated increase in a patient's diabetes risk is 5% for an increase in age by 10 years. The NLM tells us how confident we can be that aging the patient by 5 years will result in a 2.5% increase in risk.

4.2.1 Computation and interpretation

Linear reference function: A natural choice for the linear reference function is the secant intersecting both points of the forward difference (see Fig. 2). The secant for a

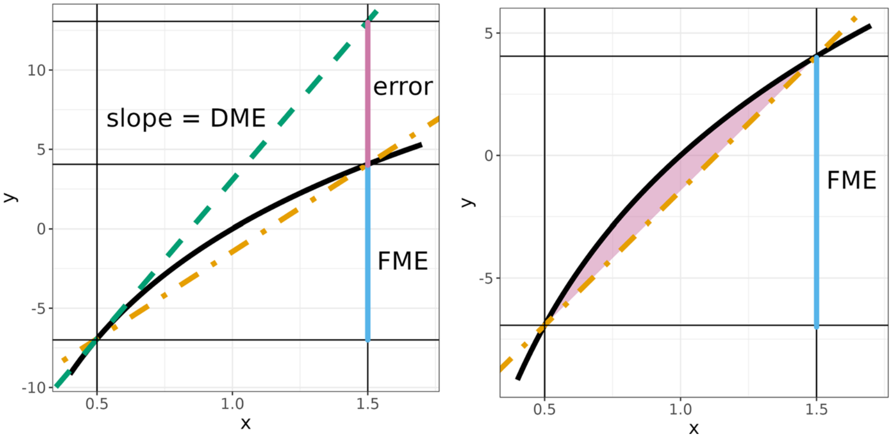


Fig. 2 Illustration of a univariate FME for $h = 1$ and a comparison to the corresponding DME. **Left**: The prediction function is black-colored. The DME is given by the slope of the tangent (green, dashed) at the point of interest (x = 0.5). The interpretation of the DME corresponds to the evaluation of the tangent value at x = 1.5, which is subject to an error (purple) almost as large as the actual change in prediction. The FME (blue) equals the change in prediction along the secant (orange, dotdashed) through the prediction at x = 0.5 and at x = 1.5. **Right**: The deviation between the prediction function (black) and linear secant (orange, dotdashed) can be quantified via the purple area. For the NLM, we put this integral in relation to the integral of the area between the prediction function and the mean prediction (Color figure online)

multivariate FME corresponds to:

$$g_{\boldsymbol{x},\boldsymbol{h}_S}(t) = \begin{pmatrix} x_1 + t \cdot h_1 \\ \vdots \\ x_s + t \cdot h_s \\ \vdots \\ x_p \\ \widehat{f}(\boldsymbol{x}) + t \cdot \mathrm{FME}_{\boldsymbol{x},\boldsymbol{h}_S} \end{pmatrix}$$

The multivariate secant considers equally proportional changes in all features. Figure 3 visualizes the discrepancy between the prediction function and the secant along a two-dimensional FME. If the NLM indicates linearity, we can infer that if *all* individual feature changes are multiplied by a scalar $t \in [0, 1]$, the FME would change by t as well.

Definition of the NLM: Comparing the prediction function against the linear reference function along the FME requires a normalized metric that indicates the degree of similarity between functions or sets of points. Established metrics in geometry include the Hausdorff (Belogay et al. 1997) and Fréchet (Alt and Godau 1995) distances. Another option is to integrate the absolute or squared deviation between both functions. These approaches have the common disadvantage of not being normalized, i.e., the degree of non-linearity is scale-dependent.

Molnar et al. (2020) compare non-linear function segments against linear models via the coefficient of determination R^2. In this case, R^2 indicates how well the linear

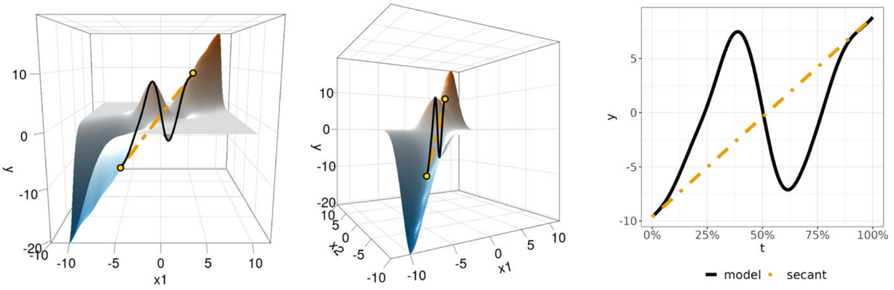


Fig. 3 A non-linear prediction function, the path along its surface, and the corresponding secant along a two-dimensional FME from point $(-5, -5)$ to point $(5, 5)$. The right plot depicts the parameterization in terms of t as the percentage of the step size $\boldsymbol{h}_S$. This type of parameterization and visualization is possible for any dimensionality of $\boldsymbol{h}_S$

reference function is able to explain the non-linear prediction function compared to the most uninformative baseline model, i.e., one that always predicts the prediction function through its mean value. As we do not have observed data points along the forward difference, points would need to be obtained through (Quasi-)Monte-Carlo sampling, whose error rates heavily depend on the number of sampled points. As both the FME and the linear reference function are evaluated along the same single path across the feature space, their deviation can be formulated as a line integral. Hence, we are able to extend the concept of R^2 to continuous integrals, comparing the integral of the squared deviation between the prediction function and the secant, and the integral of the squared deviation between the prediction function and its mean value. The line integral is univariate and can be numerically approximated with various techniques such as Gaussian quadrature.

The parametrization of the path through the feature space is given by $\gamma : [0, 1] \mapsto \mathscr{X}$, where $\gamma(0) = \boldsymbol{x}$ and $\gamma(1) = (\boldsymbol{x}_S + \boldsymbol{h}_S, \boldsymbol{x}_{-S})$. The line integral of the squared deviation between prediction function and secant along the forward difference corresponds to:

$$(\mathrm{I}) = \int_0^1 \left(\widehat{f}(\gamma(t)) - g_{\boldsymbol{x},\boldsymbol{h}_S}(t)\right)^2 \left|\left|\frac{\partial \gamma(t)}{\partial t}\right|\right|_2 dt$$

with

$$\gamma(t) = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} + t \cdot \begin{pmatrix} h_1 \\ \vdots \\ h_s \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad , \quad t \in [0, 1]$$

and

$$\left|\left|\frac{\partial \gamma(t)}{\partial t}\right|\right|_2 = \sqrt{h_1^2 + \cdots + h_s^2}$$

The integral of the squared deviation between the prediction function and the mean prediction is used as a baseline. The mean prediction is given by the integral of the prediction function along the forward difference, divided by the length of the path:

$$\begin{aligned}\overline{\widehat{f}(t)} &= \frac{\int_0^1 \widehat{f}(\gamma(t)) \left|\left|\frac{\partial \gamma(t)}{\partial t}\right|\right|_2 dt}{\int_0^1 \left|\left|\frac{\partial \gamma(t)}{\partial t}\right|\right|_2 dt} \\ &= \int_0^1 \widehat{f}(\gamma(t))\, dt \\ \text{(II)} &= \int_0^1 \left(\widehat{f}(\gamma(t)) - \overline{\widehat{f}(t)}\right)^2 \left|\left|\frac{\partial \gamma(t)}{\partial t}\right|\right|_2 dt\end{aligned}$$

The $\text{NLM}_{\boldsymbol{x},\boldsymbol{h}_S}$ is defined as:

$$\text{NLM}_{\boldsymbol{x},\boldsymbol{h}_S} = 1 - \frac{\text{(I)}}{\text{(II)}}$$

Interpretation: The NLM has an upper limit of 1 and indicates how well the secant can explain the prediction function, compared to the baseline model of using the mean prediction. For a value of 1, the prediction function is equivalent to the secant (perfect linearity). A lower value indicates increasing non-linearity of the prediction function. For negative values, the mean prediction better predicts values on the prediction function than the secant (severe non-linearity). We suggest to use 0 as a hard bound to indicate non-linearity and values on the interval]0, 1[as an optional soft bound.
Advantages of the NLM: Given only univariate changes in feature values, we may visually assess the non-linearity of the feature effect with an ICE curve (see Sect. 5). However, the NLM quantifies non-linearity in a single metric. For one, this facilitates interpretations: for instance, in Fig. 13, the average NLM correctly diagnoses linear effects of the features x_4 and x_5 in Friedman's regression problem. Second, this information can be further utilized in an informative summary output of the prediction function: in Sect. 4.3, we estimate feature effects for population subgroups where individual NLM values can be averaged to describe average non-linearities within subgroups. For bivariate feature changes, the NLM greatly simplifies non-linearity assessments: as an example, consider Fig. 12 where the sinus curve's point of inflection for the interaction of x_1 and x_2 in Friedman's regression problem can be detected with NLM values. Lastly, given changes in more than two features, visual interpretation techniques such as the ICE and PD are not applicable at all. As opposed to this, the NLM is defined in arbitrary dimensions and can be used for feature changes of any dimensionality (see Fig. 20 for an example with a trivariate feature change).

4.2.2 Selecting step sizes and linear trust region

The step size is determined both by the question that is being addressed and the scale of the feature at training time. In many cases, an interpretable or intuitive step size is preferable. For instance, body weight tends to be expressed in kilograms, thus making 1 kg (as opposed to 1 g) a natural increment. Contextual information, too, dictates step sizes. For instance, a 1 kg difference in body weight might not elicit many physiological changes. One might suspect, for instance, a 5 kg difference to elicit noticeable changes and to provide an actionable model interpretation, where the patient can be advised to lose weight if the model predicts a favorable outcome of that action. If a natural unit or contextual information is not available, the units recorded in the data set make a reasonable default step size. This also links back to the natural interpretation of LMs, whose beta coefficients indicate the change in predicted outcome due to a unit change in the feature value.

Dispersion-based step sizes: Without contextual information, dispersion-based measures such as one standard deviation can also be used as step sizes (Mize et al. 2019). Other options include, e.g., percentages of the interquartile range (IQR) or the mean / median absolute deviation. Furthermore, we can compute and visualize FME and NLM distributions for various step sizes or step size combinations for multivariate FMEs (see Fig. 18 for an example).

Local linear trust region (LLTR): In selected applications it might be of interest to have confidence in the linearity of FMEs, which can be ensured with an NLM threshold. Figure 4 visualizes an example by Molnar (2022) where LIME (see Sect. 5) fails to accurately explain the black box prediction for a data point depending on the chosen kernel width. We wish to explain the predictions of the black box model (black line) for a single data point (black dot). For kernel widths 0.75 or 2, the local surrogate indicates no or a positive effect of x on the predicted target, while the actual effect is negative. In contrast, the FME can be used to compute the exact feature effect where the NLM provides an LLTR (visualized by the orange arrows). In this example, traversing the black box model from the black dot along each arrow is associated with an NLM ≥ 0.9, i.e., an approximately linear FME. Which NLM threshold indicates linearity is debatable. For this paper, we choose a very high threshold of 0.9 to leave a margin of safety. The right plot visualizes FME and NLM pairs for each step of the LLTR. Steps that cannot be included in the LLTR are greyed out. An LLTR for multivariate steps is visualized in Fig. 17.

Step sizes and model extrapolations: The step size cannot vary indefinitely without risking model extrapolations. Furthermore, when using non-training data or training data in low-density regions to compute FMEs, we are at risk of the model extrapolating without actually traversing the feature space. If the points $\boldsymbol{x}$ or $(\boldsymbol{x}_S + \boldsymbol{h}_S, \boldsymbol{x}_{-S})$ are classified as extrapolation points (EPs), the FME should be interpreted with caution or the observation be excluded from the analysis.

Fig. 5 demonstrates the perils of model extrapolations when using FMEs. We draw points of a single feature x from a uniform distribution on the interval $[-5, 5]$. The target is generated as $y = x^2 + \epsilon$, where ϵ is drawn from $N(0, 1)$. A random forest is trained to predict y given x. All points $x \notin [-5, 5]$ are located outside the range of the training data and can be considered EPs. We compute FMEs with a step size of 1.

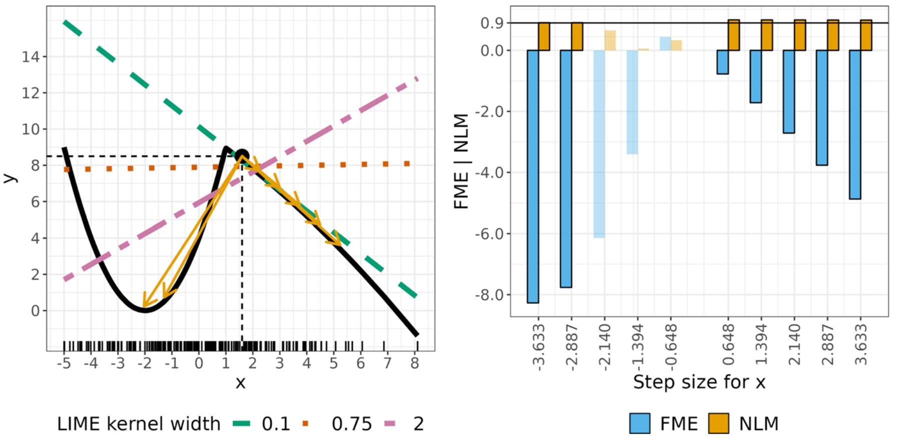


Fig. 4 **Left:** Explaining a single local prediction at the black dot (x = 1.6, $\widehat{f}(x) = 8.5$). The black box model predictions are given by the black line. Local surrogate explanations via LIME differ considerably depending on the chosen kernel width (straight lines, kernel width indicated by shape and color). In contrast, the FME always represents an exact forward difference between the black dot and points on the prediction function (where the secant is visualized by the orange arrows). The step sizes associated with the arrows represent an exemplary LLTR of FMEs for which the NLM ≥ 0.9 (approximate linearity). **Right:** Visualization of LLTR with pairs of FME and NLM for each explored step. Step sizes with an NLM < 0.9 are greyed out (Color figure online)

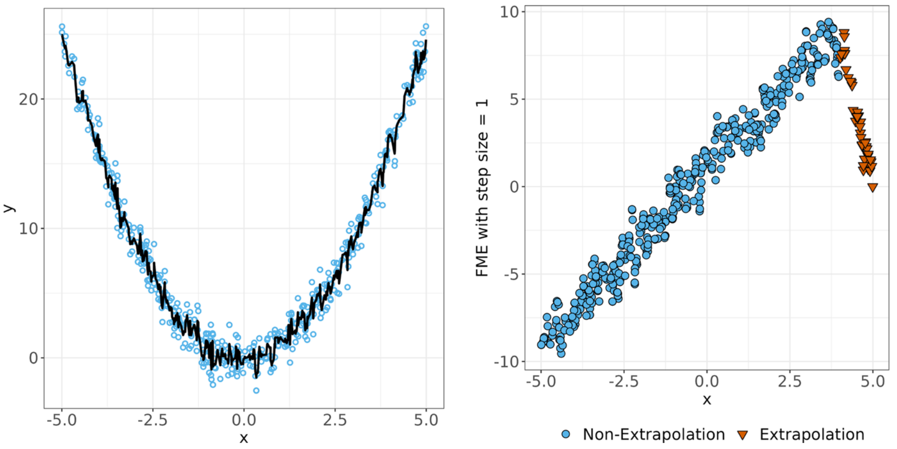


Fig. 5 **Left**: A random forest is trained on a single feature x with a quadratic effect on the target. The training space corresponds to the interval $[-5, 5]$. **Right**: We compute an FME with a step size of 1 for each observation. After moving 1 unit in x direction, points with $x > 4$ are considered EPs (red triangles). The random forest extrapolates and predicts unreliably in this area of the feature space. The resulting FMEs are irregular and should not be used for interpretation purposes (Color figure online)

By implication, all FMEs with $x > 4$ are based on model extrapolations. FMEs based on model extrapolations exhibit a considerably different pattern and should not be used for interpretation purposes, as they convey an incorrect impression of the feature effect of x.

4.3 Regional feature effects with conditional average marginal effects

It is desirable to summarize the feature effect in a single metric, similarly to the parameter-focused interpretation of LMs. For instance, one is often interested in the expected FME (for the entire feature space), which can be estimated via the AME. However, averaging heterogeneous FMEs to the AME is not globally representative of non-linear prediction functions such as the ones created by ML algorithms. A heterogeneous distribution of FMEs requires a more local evaluation. As opposed to conditioning on feature values in the case of MERs (local), we further suggest to condition on specific feature subspaces (regional). The cAME is an estimate of the expected FME for the random vector $\boldsymbol{X}_{[\,]}$ with a restricted sample space $\mathscr{X}_{[\,]}$. It is computed for a subsample of observations $\mathscr{D}_{[\,]}$ sampled from $\mathscr{X}_{[\,]}$:

$$\begin{aligned} \mathrm{cAME}_{\mathscr{D}_{[]},\boldsymbol{h}_S} &= \mathbb{E}_{\boldsymbol{X}_{[]}}\left[\widehat{\mathrm{FME}_{\boldsymbol{X}_{[]}}}_{,\boldsymbol{h}_S}\right] \\ &= \frac{1}{n_{[]}} \sum_{i:\boldsymbol{x}^{(i)} \in \mathscr{D}_{[]}} \left[\widehat{f}\left(\boldsymbol{x}_S^{(i)} + \boldsymbol{h}_S, \boldsymbol{x}_{-S}^{(i)}\right) - \widehat{f}\left(\boldsymbol{x}^{(i)}\right)\right] \\ &\text{with} \quad n_{[]} = |\mathscr{D}_{[]}| \end{aligned} \tag{2}$$

A population subgroup $\mathscr{X}_{[\,]}$ corresponds to a subspace of the feature space $\mathscr{X}$, e.g., a range of health characteristics of patients with a certain predisposition of developing a disease. The subsample $\mathscr{D}_{[\,]}$ consists of data that were drawn from this subspace, e.g., patients with said predisposition that partook in a study. Note that in our case, we are looking for subgroups with homogeneous effects on the model prediction, e.g., patients for whom increasing their age has similar effects on the predicted disease risk. Even though such population subgroups might exist (in many cases they may not), the model fit fundamentally determines whether we can find subgroups with homogeneous effects for the trained model.

4.3.1 Desiderata for finding subgroups

Note that Eq. (2) is defined in general terms, conditional on an arbitrary subspace $\mathscr{X}_{[\,]}$. We can arbitrarily partition the feature space, determine corresponding subsets of observed data, and run the estimator in Eq. (2) for each subsample to estimate expected conditional FMEs. However, recall that our goal is to find accurate descriptors of feature effects for the trained model across the feature space. Therefore, we formulate multiple desiderata for these subspaces and the corresponding subsamples (hereafter referred to as subgroups):

- **Within-group effect homogeneity:** FME variance inside subgroups shall be minimized.
- **Between-group effect heterogeneity:** cAMEs of subgroups shall be heterogeneous.
- **Full segmentation:** The data shall be fully segmented into subgroups.
- **Non-congruence:** Subgroups shall not overlap with each other.
- **Confidence:** Larger subgroups are preferred over smaller subgroups.

- **Stability:** Subgroups shall be stable w.r.t. variations in the data.

Evidently, certain desiderata are difficult to meet. For instance, we can strive to minimize FME variance within a single subgroup, but this might increase FME variance within other subgroups.

Note that the philosophy of regional or semi-global feature effects somewhat deviates from our previous philosophy of obtaining simple and stable local model explanations. Finding subgroups with more homogeneous local explanations by modeling FME patterns necessitates some sort of approximation. In the following section, we model FME patterns with decision trees and discuss the upsides and downsides of this approach.

4.3.2 Estimation using decision trees

Decision tree learning is an ideal scheme to partition the entire feature space into mutually exclusive subspaces, thus finding population subgroups. Growing a tree by global optimization poses considerable computational difficulties and corresponds to an NP-complete problem (Norouzi et al. 2015). Recent developments in computer science and engineering can be explored to revisit global decision tree optimization from a different perspective, e.g., Bertsimas and Dunn (2017) explore mixed-integer optimization to find globally optimal decision trees. To reduce computational complexity, the established way (which is also commonly available in many software implementations) is through recursive partitioning (RP), optimizing an objective function in a greedy way for each tree node.

Over the last decades, a large variety of RP methods has been proposed (Loh 2014), with no gold standard having crystallized to date. In principle, any RP method that is able to process continuous targets can be used to find subgroups, e.g., classification and regression trees (CART) (Breiman et al. 1984; Hastie et al. 2001), which is one of the most popular approaches. Trees have been demonstrated to be notoriously unstable w.r.t. perturbations in input data (Zhou et al. 2023; Last et al. 2002). Tree ensembles, such as random forests (Breiman 2001a), reduce variance but lose interpretability as a single tree structure. Exchanging splits along a single path results in structurally different but logically equivalent trees (Turney 1995). It follows that two structurally very distinct trees can create the same or similar subspaces. We are therefore not interested in the structure of the tree itself, but in the subgroups it induces.

Stabilization of RP: As formulated earlier, we strive to find subgroups that are stable w.r.t. variations in the data. For RP, one should therefore strive to stabilize splits. A branch of RP methods incorporates statistical theory into the split procedure. Variants include conditional inference trees (CTREE) (Hothorn et al. 2006), which use a permutation test to find statistically significant splits; model-based recursive partitioning (MOB) (Zeileis et al. 2008), which fits node models and tests the instability of the model parameters w.r.t. partitioning the data; or approximation trees (Zhou et al. 2023), which generate artificially created samples for significance testing of tree splits. Seibold et al. (2016) use MOB to find patient subgroups with similar treatment effects in a medical context. Furthermore, we can assess the stability of feature and split point selection for arbitrary tree models by resampling the training data and retraining the tree (Philipp et al. 2016). The variance and instability of decision trees partly stems from binary splits, as a decision higher up cascades through the entire tree and results in different splits lower down the tree (Hastie et al. 2001). Using multiway trees, which also partition the entire feature space, would therefore improve stability. However, multiway splits are associated with a considerable increase in computational complexity and are therefore often discarded in favor of binary splitting (Zeileis et al. 2008). For the remainder of the paper, we use CTREE to find subgroups and compute cAMEs.

4.3.3 Confidence intervals for the cAME and cANLM

Given estimates of the expected conditional FME, it is desirable to estimate the expected conditional NLM for the corresponding subspaces as well. Analogously to the AME, we can compute an average NLM (ANLM) by globally averaging NLMs and a conditional ANLM (cANLM) by averaging NLMs within a subgroup. The cANLM gives us an estimate of the expected non-linearity of the prediction function for the given movements along the feature space, conditional on a feature subspace.

A lower standard deviation (SD) of FMEs and NLM values increases confidence in our estimates and vice versa, and a larger number of observations increases confidence in our estimates and vice versa. Although we do not specify a distribution of the underlying FMEs or NLMs, constructing a confidence interval (CI) is possible via the central limit theorem. As the cAME and cANLM are sample averages of all FMEs and NLMs for each subgroup, we can construct a t-statistic (as the SD is estimated) for large sample sizes. Given a subgroup $\mathscr{D}_{[\,]}$ that contains $n_{[\,]}$ observations, mean ($\text{cAME}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S}$ and $\text{cANLM}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S}$) and SD ($\text{SD}_{\text{FME},\,[\,]}$ and $\text{SD}_{\text{NLM},\,[\,]}$) values, the confidence level α, and the values of the t-statistic with $n_{[\,]} - 1$ degrees of freedom at the $1 - \frac{\alpha}{2}$ percentile ($t_{1-\frac{\alpha}{2},\, n_{[\,]}-1}$), the CIs correspond to:

$$\text{CI}_{\text{cAME},\,1-\alpha} = \Bigg[\text{cAME}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S} - t_{1-\frac{\alpha}{2},\, n_{[\,]}-1} \frac{\text{SD}_{\text{FME},\,[\,]}}{\sqrt{n_{[\,]}}}\,, \quad \text{cAME}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S} + t_{1-\frac{\alpha}{2},\, n_{[\,]}-1} \frac{\text{SD}_{\text{FME},\,[\,]}}{\sqrt{n_{[\,]}}}\Bigg]$$

$$\text{CI}_{\text{cANLM},\,1-\alpha} = \Bigg[\text{cANLM}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S} - t_{1-\frac{\alpha}{2},\, n_{[\,]}-1} \frac{\text{SD}_{\text{NLM},\,[\,]}}{\sqrt{n_{[\,]}}}\,, \quad \text{cANLM}_{\mathscr{D}_{[\,]},\boldsymbol{h}_S} + t_{1-\frac{\alpha}{2},\, n_{[\,]}-1} \frac{\text{SD}_{\text{NLM},\,[\,]}}{\sqrt{n_{[\,]}}}\Bigg]$$

One option to ensure that the lower sample size threshold for CIs is valid is to specify a minimum size for each subgroup, e.g., in the case of RP, not growing the tree too large.

5 Related work

5.1 Statistics and applied fields

MEs have been discussed extensively in the literature on statistics and statistical software, e.g., by Ai and Norton (2003), Greene (2012), Norton et al. (2019), or Mullahy (2017). The `margins` command is a part of Stata (StataCorp. 2023) and was originally implemented by Bartus (2005). A brief description of the `margins` command is given by Williams (2012). Leeper (2018) provides an overview on DMEs and their variations as well as a port of Stata's functionality to R. The R package `marginaleffects` (Arel-Bundock 2023) supports various variants of MEs including FMEs. Ramsey and Bergtold (2021) compute an ME for a single-hidden-layer feed-forward backpropagation artificial neural network by demonstrating its interpretation is equivalent to a logistic regression model with a flexible index function. Zhao et al. (2020) apply model-agnostic DMEs to ML models in the context of analyzing travel behavior. Furthermore, they mention the unsuitability of derivatives for tree-based prediction functions such as random forests.

Mize et al. (2019) provide a test framework for cross-model differences of MEs. They refer to an ME based on a forward difference as a discrete change and to the corresponding AMEs as average discrete changes. Gelman and Pardoe (2007) propose the predictive effect as a local feature effect measure. The predictive effect is a univariate forward difference, divided by the change in feature values (i.e., the step size). This differentiates it from the FME which is also defined for multivariate feature changes and which is not divided by the step size, i.e., it provides a change in prediction as opposed to a rate of change. Furthermore, the authors propose an average predictive effect that corresponds to the average of multiple predictive effects that were measured at distinct feature values and model parameters. It is a generalization of the AME that may be estimated with artificially created data points (as opposed to the sample at hand) and incorporates model comparisons (measured with different model parameters).

5.2 Interpretable machine learning

The most commonly used techniques to determine feature effects include the individual conditional expectation (ICE) (Goldstein et al. 2015), the partial dependence (PD) (Friedman 2001), accumulated local effects (ALE) (Apley and Zhu 2020), Shapley values (Štrumbelj and Kononenko 2014), Shapley additive explanations (SHAP) (Lundberg and Lee 2017), local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), and counterfactual explanations (Wachter et al. 2018). Counterfactual explanations indicate the smallest necessary change in feature values to receive the desired prediction and represent the counterpart to MEs. Goldstein et al. (2015) propose derivative ICE (d-ICE) plots to detect interactions. The d-ICE is a univariate ICE where the numeric derivative w.r.t. the feature of interest is computed pointwise after a smoothing procedure. Symbolic derivatives are commonly used to determine the importance of features for neural networks (Ancona et al. 2018). While FMEs provide interpretations in terms of prediction *changes*, most methods provide an interpretation in terms of prediction *levels*. LIME is an alternative option that returns interpretable parameters (i.e., rates of change in prediction) of a local surrogate model. LIME, and to a lesser extent SHAP, have been demonstrated to provide unreliable interpretations in some cases (Slack et al. 2020). Furthermore, many techniques in IML are interpreted visually (e.g., ICEs, the PD, ALE plots) and are therefore limited to feature value changes in at most two dimensions. FMEs are not limited by the dimension of the intervention in feature values, as any change in feature values—regardless of its dimensionality—always results in a single FME.

5.2.1 Relation between forward marginal effects, the individual conditional expectation, and partial dependence

Given a data point $\boldsymbol{x}$, the ICE of a feature set S corresponds to the prediction as a function of substituted values $\boldsymbol{x}_S^*$ where $\boldsymbol{x}_{-S}$ is kept constant:

$$\text{ICE}_{\boldsymbol{x},S}(\boldsymbol{x}_S^*) = \widehat{f}(\boldsymbol{x}_S^*, \boldsymbol{x}_{-S})$$

The PD on a feature set S corresponds to the expectation of $\widehat{f}(\boldsymbol{X})$ w.r.t. the marginal distribution of $\boldsymbol{X}_{-S}$. It is estimated via Monte-Carlo integration where the draws $\boldsymbol{x}_{-S}$ correspond to the sample values:

$$\widehat{\text{PD}}_{\mathscr{D},S}(\boldsymbol{x}_S) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}\left(\boldsymbol{x}_S, \boldsymbol{x}_{-S}{}^{(i)}\right)$$

We can visually demonstrate that in the univariate case, the FME is equivalent to the vertical difference between two points on an ICE curve. However, the AME is only equivalent to the vertical difference between two points on the PD curve for linear prediction functions (see Fig. 6). We generalize this result to the multivariate FME and ICE, as well as the multivariate forward AME and PD (see Theorem 3 and Theorem 4 in Appendix B.2). Visually assessing changes in prediction due to a change in feature values is difficult to impossible in more than two dimensions. High-dimensional feature value changes therefore pose a natural advantage for FMEs over techniques such as the ICE, PD, or ALE, which are mainly interpreted visually.

5.2.2 Comparison to LIME

LIME—one of the most popular model-agnostic feature effect methods—resembles the interpretation given by an FME. It also serves as a local technique, explaining the model for a single observation. LIME samples instances, predicts, and weights the predictions by the instances' proximity to the instance of interest using a kernel function. Afterwards, an interpretable surrogate model is trained on the weighted predictions. The authors choose a sparse linear model, whose beta coefficients provide an interpretation similar to the FME.

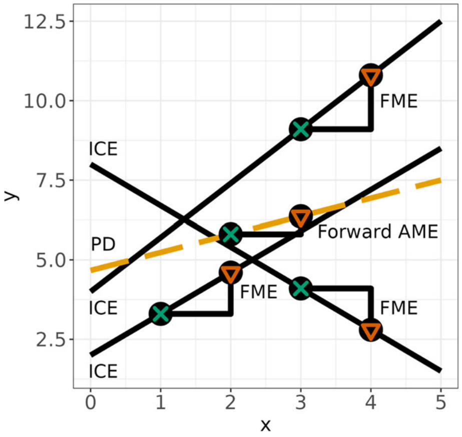


Fig. 6 Three ICE curves are black-colored. The PD is the average of all ICE curves (orange, dashed). For each ICE curve, we have a single observation, visualized by the corresponding green x-shaped point. We compute the FME at each observation with a step size of 1, which results in the corresponding red triangle-shaped point. The FMEs are equivalent to the vertical difference between two points on the ICE curves. If the prediction function is linear in the feature of interest, the average of all FMEs is equivalent to the vertical difference between two points on the PD (Color figure online)

But there is a fundamental difference between both approaches. The FME directly works on the prediction function, while LIME trains a local surrogate model. The latter is therefore affected by an additional layer of complexity and uncertainty. The authors suggest to use LASSO regression, which requires choosing a regularization constant. Furthermore, one must select a similarity kernel defined on a distance function with a width parameter which has tremendous effects on the resulting model explanation (see Fig. 4 for an example). The model interpretation is therefore fundamentally determined by multiple parameters. Furthermore, certain surrogate models are incapable of explaining certain model behaviors and may potentially mislead the practitioner to believe the interpretation (Ribeiro et al. 2016). A linear surrogate model may not be able to describe extreme non-linearities of the prediction function, even within a single locality of the feature space. In contrast, the only parameters for the FME are the features and the step sizes. Without question, the choice of parameters for FMEs also significantly affects the interpretation. However, we argue that their impact is much clearer than in LIME, e.g., a change in a feature such as age is much more meaningful than a different width parameter in LIME. In fact, we argue that the motivation behind both approaches is fundamentally different. For FMEs, we start with a meaningful interpretation concept in mind, e.g., we may be interested in the combined effects of increasing age and weight on the disease risk. For LIME, we start with a single observation, trying to distill the black box model behavior within this specific locality into a surrogate model.

In addition to the sensitivity of results regarding parameter choices, LIME is notoriously unstable even with fixed parameters. Zhou et al. (2021) note that repeated runs using the same explanation algorithm on the same model for the same observation results in different model explanations, and they suggest significance testing as a remedy. In contrast, FMEs with fixed parameters are deterministic.

As noted above, the authors of LIME mention that the faithfulness of the local surrogate may be diminished by extreme non-linearities of the model, even within the locality of the instance of interest. This exact same critique holds for the FME (see Sect. 4.2). Hence, we introduce the NLM, which essentially corresponds to a measure of faithfulness of the FME and whose concept can potentially be used for other methods as well. One could also use the coefficient of determination R^2 to measure the goodness-of-fit of the linear surrogate to the pseudo sample in LIME. However, we argue that the goodness-of-fit to a highly uncertain pseudo sample is a questionable way of measuring an explanation's faithfulness.

Furthermore, the authors of LIME note that insights into the global workings of the model may be gained by evaluating multiple local explanations. As there usually are time constraints so that not all instances can be evaluated, an algorithm suggests a subset of representative instances. Although this approach avoids the issue of misrepresenting global effects by averaging local explanations, it also misses the opportunity to provide meaningful regional explanations. This is where the cAME comes into play. It is motivated by the goal to aggregate local interpretations while staying faithful to the underlying predictive model. Note that a subset of representative instances—as suggested by Ribeiro et al. (2016)—can also be used to compute representative FMEs.

5.3 Sensitivity analysis

The goal of sensitivity analysis (SA) is to determine how uncertainty in the model output can be attributed to uncertainty in the model input, i.e., determining the importance of input variables (Saltelli et al. 2008). Techniques based on FDs are common in SA (Razavi et al. 2021). The numeric derivative of the function to be evaluated w.r.t. an input variable serves as the natural definition of local importance in SA. The elementary effect (EE) was first introduced as part of the Morris method (Morris 1991) as a screening tool for important inputs. The EE corresponds to a univariate forward difference quotient with variable step sizes, i.e., it is a generalization of the derivative. Variogram-based methods analyze forward differences computed at numerous pairs of points across the feature space (Razavi and Gupta 2016). Derivative-based global sensitivity measures (Sobol and Kucherenko 2010) provide a global feature importance metric by averaging derivatives at points obtained via random or quasi-random sampling.

6 Simulations

Here, we present multiple simulation scenarios to highlight the workings and interplay of FMEs, the NLM, and the cAME. In all following sections, we use Simpson's 3/8 rule for the computation of the NLM and CTREE to find subgroups.

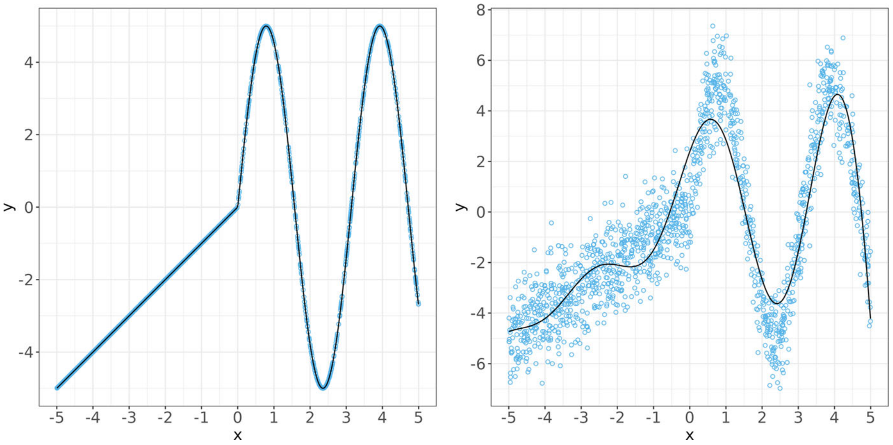


Fig. 7 The target is determined by a single feature x. On the interval $[-5, 0[$ there is a linear feature effect. On the interval $[0, 5]$ the functional relationship consists of a transformed sine wave. We first use the DGP, then add random noise on top of the data and train an SVM (Color figure online)

6.1 Univariate data without noise

We start with a univariate scenario without random noise and work directly with the data generating process (DGP). This way, we can evaluate how introducing noise affects the information gained from FMEs in the subsequent simulation. We simulate a single feature x, uniformly distributed on $[-5, 5]$, and define f as:

$$f(x) = \begin{cases} x & x < 0 \\ 5\sin(2x) & x \geq 0 \end{cases}$$

The data are visualized in Fig. 7. An FME with step size $h = 2$ is computed for each observation. We use CTREE on the FMEs to find subgroups. Subsequently, all observations' NLM values are averaged to cANLM values on the subspaces of the cAMEs. Our computations are visualized in Fig. 8. Vertical lines indicate tree splits, and corresponding FME or NLM subgroup averages are indicated by horizontal lines. In the univariate case, we see a direct relationship between the shape of the DGP and the FMEs and NLM values. The NLM has ramifications on the interpretation of the FMEs. For instance, for $x = -3$, increasing x by 2 units increases the predicted target value by 2 units, and we can conclude that the same holds proportionally for feature value changes of smaller magnitudes, e.g., a change of 1 unit results in an FME of 1, etc. On the contrary, given an observation $x = 1$, the NLM indicates considerable non-linearity. For this observation, we cannot draw conclusions about FMEs with smaller step sizes than 2 units.

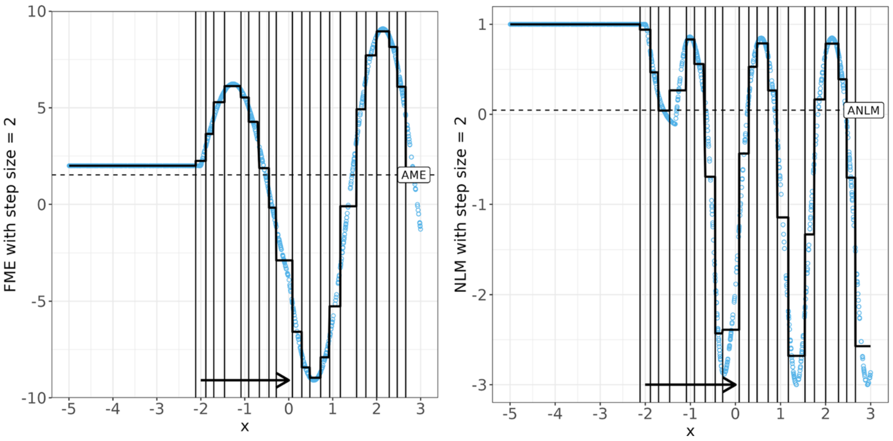


Fig. 8 **Univariate data without noise**. For each point, moving in x direction by the length of the arrow results in the FME / NLM indicated on the vertical axis. **Left**: FMEs with step size $h = 2$. A regression tree partitions the feature space into subspaces (in this case intervals) where the FMEs are most homogeneous. The horizontal lines correspond to the cAMEs. **Right**: NLM values and cANLMs for each subspace (Color figure online)

6.2 Univariate data with noise

We proceed to add random noise $\epsilon \sim N(0, 1)$ on top of the data and tune the regularization and sigma parameters of a support vector machine (SVM) with a radial basis function kernel (see Fig. 7). As we now employ a predictive model, we must avoid potential model extrapolations. The forward location of all points with $x > 3$ falls outside the range of the training data. After removing all extrapolation points, we evaluate the FMEs and NLMs of all observations with $x \in [-5, 3]$ (see Fig. 9). In this case, we can visually assess that the predictions of the SVM resemble the DGP but also factor in noise (see Fig. 7). e.g., the SVM prediction function is non-linear in linear regions of the DGP, which affects the FMEs and NLMs. This demonstrates that FMEs can only be used to explain the DGP if the model describes it accurately.

6.3 Bivariate data with univariate feature change

We next augment the univariate data with one additional feature in order to empirically evaluate the additive recovery property of the FME (see Appendix B.1). Due to potential model extrapolations, we only make use of the DGP. In the first example, the DGP corresponds to a supplementary additively linked feature x_2:

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & x_1 < 0 \\ 5\sin(2x_1) + x_2 & x_1 \geq 0 \end{cases}$$

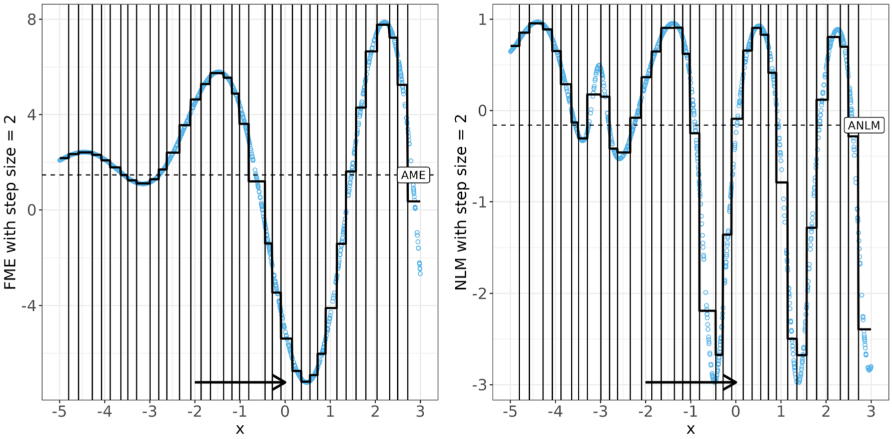


Fig. 9 **Univariate data with noise**. For each point, moving in x direction by the length of the arrow results in the FME / NLM indicated on the vertical axis. **Left**: FMEs with step size $h = 2$ and cAMEs. **Right**: NLM values and cANLMs (Color figure online)

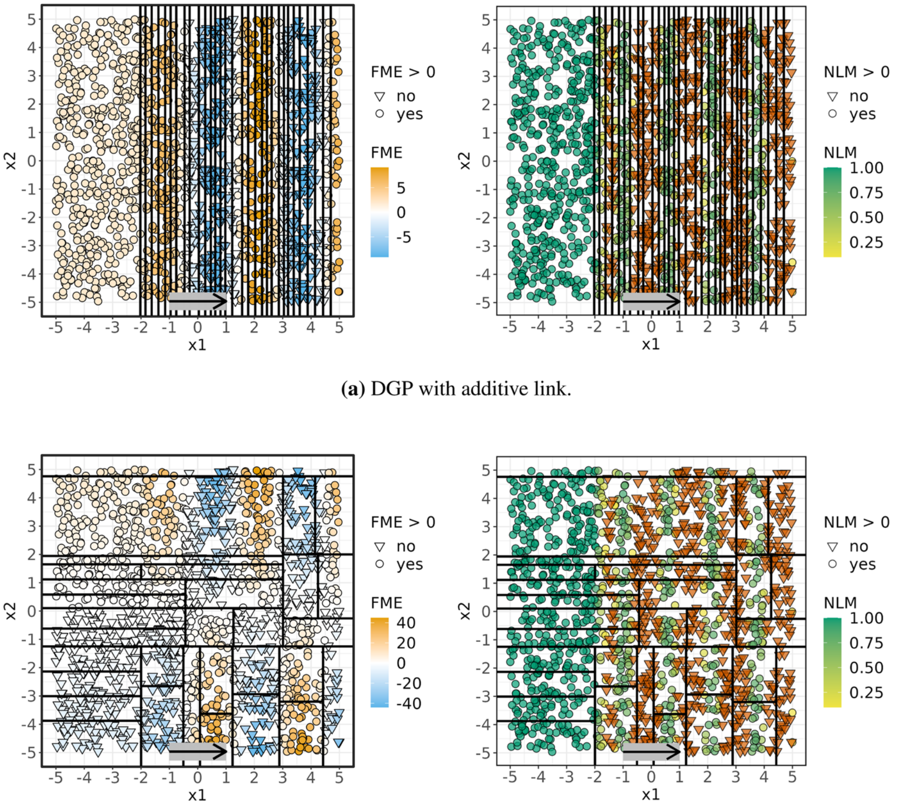
In the second example, the DGP corresponds to a supplementary multiplicatively linked feature x_2, i.e., we have a pure interaction:

$$f(x_1, x_2) = \begin{cases} x_1 \cdot x_2 & x_1 < 0 \\ 5\sin(2x_1) \cdot x_2 & x_1 \geq 0 \end{cases}$$

The FMEs and NLM values for both DGPs are given in Fig. 10. For the additive DGP, given the value of x_1, moving in x_2 direction does not influence the FMEs due to the additive recovery property. As a result, we receive the same FMEs with an additively linked feature x_2 as without it (as long as the feature change does not occur in x_2). For the multiplicative DGP, the FMEs now vary for a given x_1 value, even though the feature change only occurs in x_1. The NLM values are both affected by the presence of an additively linked and a multiplicatively linked feature x_2, even though the feature change only occurs in x_1. As opposed to the additive DGP, the cAME tree makes use of x_2 as a split variable for the multiplicative DGP.

6.4 Bivariate data with bivariate feature change

Next, we demonstrate bivariate FMEs and the corresponding NLM. We use the same DGPs as for the univariate feature change. The FMEs and NLM values are given in Fig. 11. As opposed to the univariate feature change for additively linked data, the FME values now also vary in x_2 direction for a given x_1 value due to the simultaneous change in x_2. The NLM indicates linearity for a multitude of observations, given both the additive and the multiplicative DGP. For these observations, we can infer that multiplying *both* step sizes by a value on the interval [0, 1] results in an equally proportionally reduced FME.



(a) DGP with additive link.

(b) DGP with multiplicative link.

Fig. 10 **Bivariate data and univariate feature change $\mathbf{h_1 = 2}$**. For each point, moving in x_1 direction by the length of the arrow results in the FME / NLM indicated by the color. FMEs (left) and NLM (right). Negative NLM values are red-colored (Color figure online)

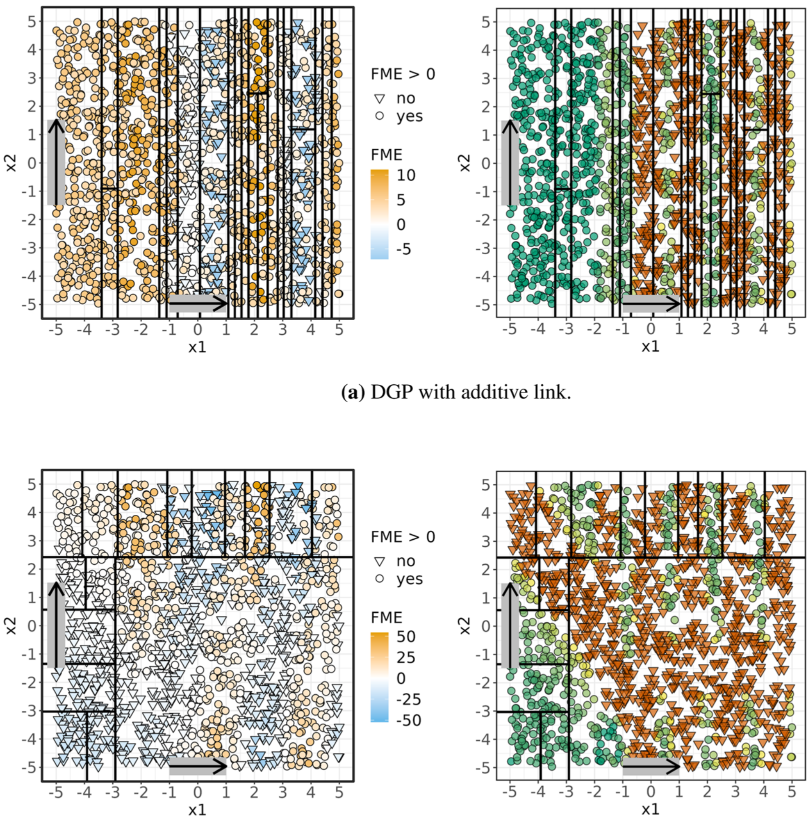
6.5 Friedman's regression problem

In the last simulation example, we demonstrate how FMEs are able to discover effects within a higher-dimensional function.[4] In Friedman's regression problem (Friedman 1991; Breiman 1996), we have 10 independent and uniformly distributed variables on the interval [0, 1]. The target is generated using the first 5 variables:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon$$

where ϵ is drawn from $N(0, \sigma)$. We simulate 1000 instances with $\sigma = 0$ and tune the regularization and sigma parameters of an SVM with a radial basis function kernel on all 10 features. Recall that our ability to conduct inference regarding the DGP depends on how well the model approximates it. In the following illustrations, we select an

[4] As our goal is to recover terms within the DGP, we refrain from computing cAMEs here.



(a) DGP with additive link.

(b) DGP with multiplicative link.

Fig. 11 **Bivariate data and bivariate feature change $\mathbf{h_1} = \mathbf{2}$ and $\mathbf{h_2} = \mathbf{3}$**. For each point, moving in x_1 and x_2 directions by the lengths of the respective arrows results in the FME / NLM indicated by the color. FMEs (left) and NLM (right). Negative NLM values are red-colored (Color figure online)

identical step size of 0.1 for each feature. As this represents roughly 10% of each feature's range, it facilitates the comparison between FMEs and the expected effect within the DGP. For instance, with a step size of 0.1 for x_5, we expect an AME of $5 \cdot 0.1 = 0.5$ if the model has a good fit. In this example, negative NLM values are set to zero (which acts as a hard bound for non-linearity) to compute the ANLM.

We first analyze the interaction pair x_1 and x_2 (see Fig. 12). For small values of either x_1 or x_2, univariate FMEs are mostly positive, while for feature values larger than 0.5, they are increasingly negative. Bivariate FMEs are largest for medium value combinations of x_1 and x_2 or large values of one feature and small values of the other. FMEs are negative for the product of x_1 and x_2 approaching 1. This, too, is expected since the sinus curve's point of inflection is located at $\frac{\pi}{2}$, and the blue area of negative FMEs roughly corresponds to $\frac{\pi}{2} = \pi x_1 x_2$, e.g., for $x_1 = x_2 \approx 0.707$. The bivariate NLM confirms our analysis by indicating strong non-linearity in said area of the sinus curve's point of inflection.

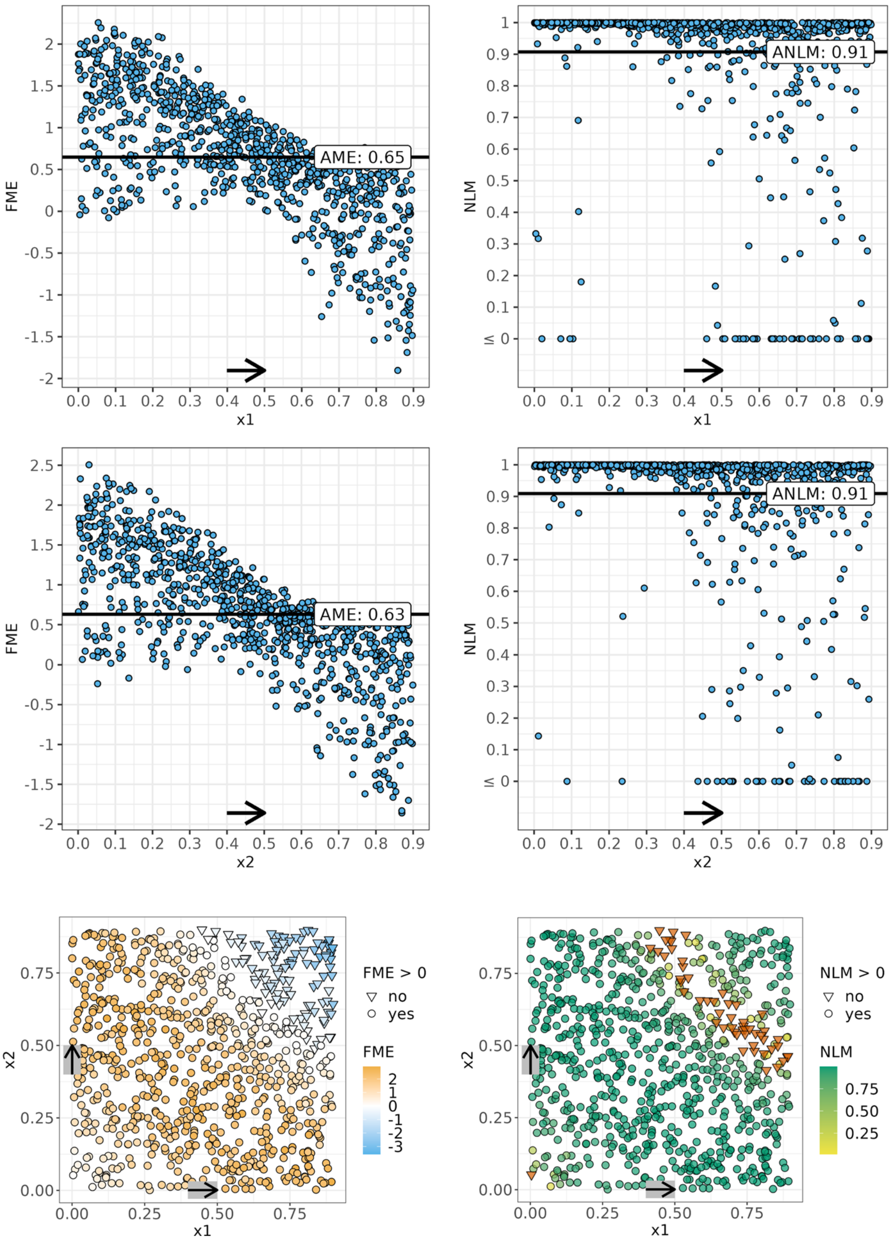


Fig. 12 Friedman's regression problem: Univariate and bivariate FMEs and NLMs for x_1 and x_2 with step sizes of 0.1 for both features. Negative NLM values are red-colored. Around the sinus curve's point of inflection, FMEs turn negative, and the NLM clearly diagnoses non-linearity (red triangles)) (Color figure online)

Next, we evaluate univariate effects of x_3, x_4, and x_5 (see Fig. 13). For x_3, we can see a linear trend of FMEs, which are mostly negative for values smaller than 0.5 and positive for values larger than 0.5. This is expected, since the effect of x_3 within the DGP is quadratic but shifted by 0.5 to the right. The NLM correctly diagnoses strong linearity for small and large values of x_3 but non-linearity for the point of inflection. Both x_4 and x_5 have positive linear effects on the target within the DGP, with the effect of x_4 being twice as large as the effect of x_5. Given the DGP, we would expect an increase of 0.1 in x_4 to have an AME of 1 (observed AME = 0.92) and an increase of 0.1 in x_5 to have an AME of 0.5 (observed AME = 0.46). FMEs reveal both linear patterns with the AMEs closely recovering expected effects and the NLMs indicating strong linearity.

Lastly, we evaluate FMEs for x_6 (see Fig. 14) which has no effect on the target within the DGP. We can see a cluster of FMEs, roughly without any correlations. The AME is approximately zero, thus accurately recovering the (non-existent) feature effect of x_6.

7 Application workflow and applied example

We now present a structured application workflow that incorporates the theory presented in the preceding sections and apply it to real data:

1. Train and tune a predictive model.
2. Based on the application context, choose evaluation points $\mathscr{D}$, the features of interest S, and the step sizes $\boldsymbol{h}_S$.
3. Check whether any $\boldsymbol{x}^{(i)}$ or $(\boldsymbol{x}_S{}^{(i)} + \boldsymbol{h}_S, \boldsymbol{x}_{-S}{}^{(i)})$ are subject to model extrapolations. See Appendix A.2 for possible options.
4. Either modify step sizes so no points are subject to model extrapolations or remove the ones that are.
5. Compute FMEs for selected observations and the chosen step sizes.
6. Optional: Compute the NLM for every computed FME.
7. Optional: Compute cAMEs by finding subgroups with homogeneous FMEs.
8. Optional: Compute cANLM values.
9. Optional: Compute CIs for cAME and cANLM.
10. Conduct local (single FMEs of interest) and (optionally) regional interpretations (cAME and cANLM).

The white wine data set (Cortez et al. 2009) consists of 4898 white wines produced in Portugal. The target is the perceived preference score of wine testers on a scale of 1-10, which we model as a continuous variable. The features consist of wine characteristics such as alcohol by volume (ABV) or the pH value. We start by tuning the regularization and sigma parameters of an SVM with a radial basis function kernel.

We first compare our results to the analysis by Goldstein et al. (2015) who train a neural network with 3 hidden units. They note that their model might be subject to performance issues and that their analysis shall only exemplify the types of interpretations ICE curves are able to generate. Model-agnostic interpretations are conditional on the trained model and can only be vaguely compared. In their analysis, the effect

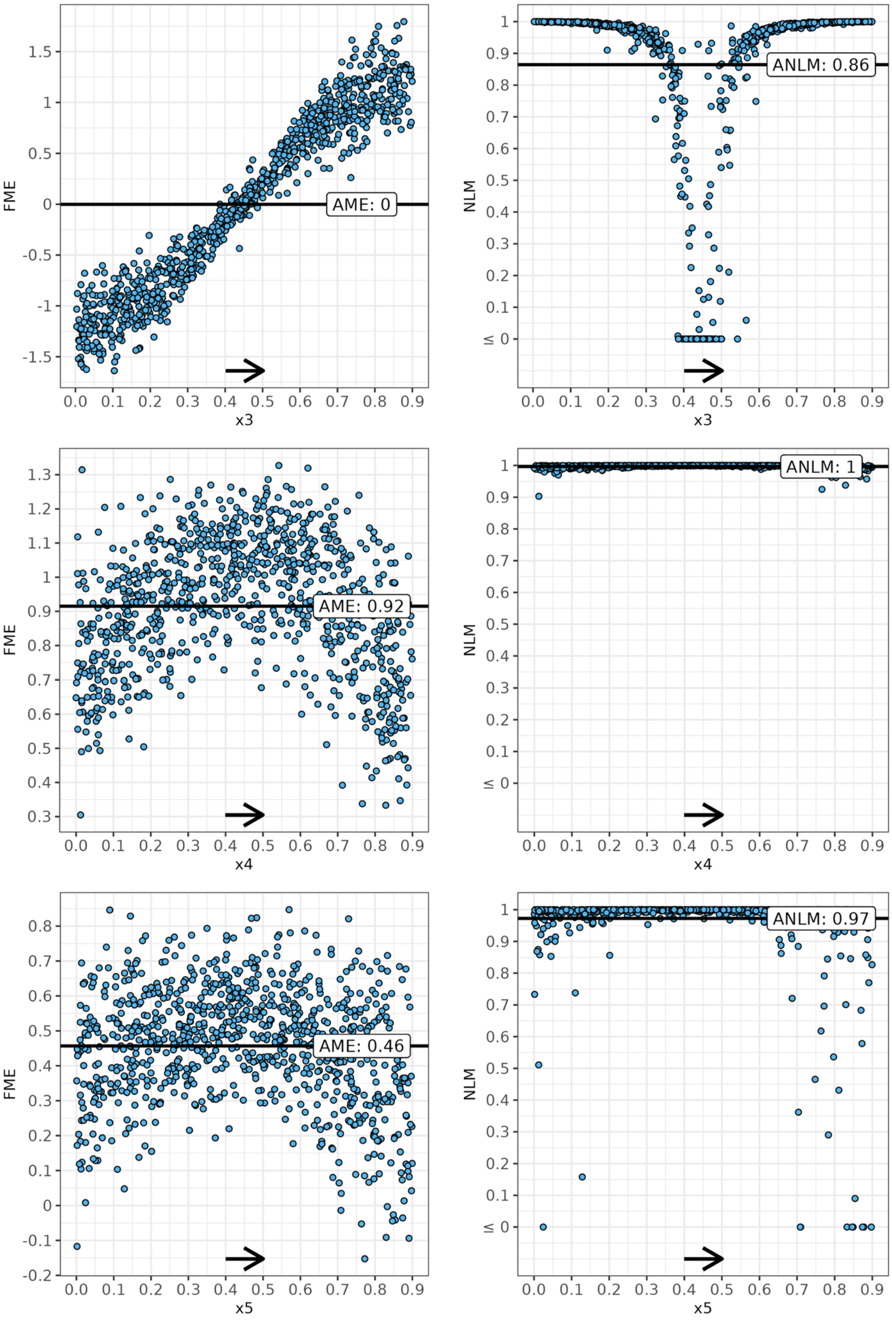


Fig. 13 **Friedman's regression problem:** Univariate FMEs and NLMs with step sizes of 0.1 for features x_3, x_4, and x_5. The NLM indicates non-linearity around the point of inflection of the quadratic effect of x_3. It indicates strong linearity for x_4 and x_5 which have linear effects on the simulated target. AMEs approximately recover the expected FME within the DGP for x_4 and x_5 (Color figure online)

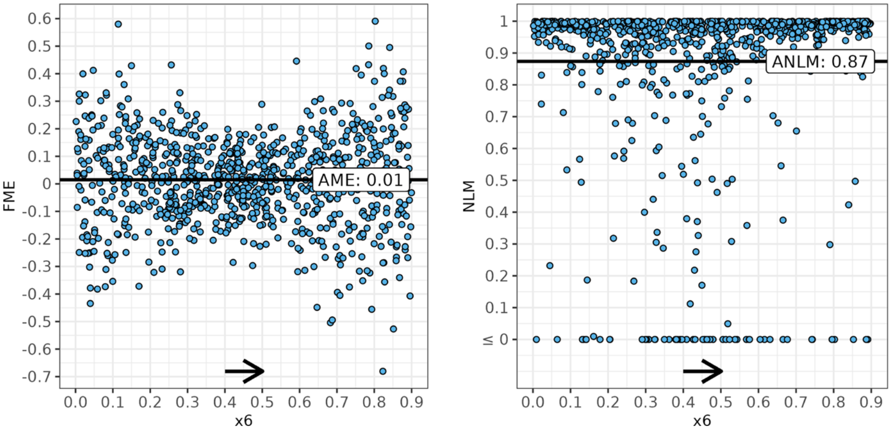


Fig. 14 **Friedman's regression problem:** Univariate FMEs and NLMs for feature x_6 with a step size of 0.1. FMEs do not exhibit any pattern, and the AME is approximately zero. This correctly diagnoses that x_6 has no effect on the simulated target (Color figure online)

of increasing the pH value on the predicted wine rating differs regarding the wine's alcohol content. In Fig. 15, we compute univariate FMEs of the pH value for a step size of 0.3 (range 2.72 to 3.82). Wines that fall outside the multivariate envelope of the training data are excluded from the analysis. The AME ≈ 0 suggests there is no global feature effect. We use CTREE to search for exactly one split and observe that a wine's alcohol content induces subgroups of a positive cAME of 0.12 (low alcohol) and a negative cAME of -0.39 (high alcohol). Resampling 500 times with 63.2% of the data results in the same split every time. This confirms our proposition that global aggregations are generally not a good descriptor of feature effects and that dividing the data into subgroups lets us discover varying cAMEs. Our methods add new insights compared to ICEs by automatically detecting the interaction between the pH value and alcohol content.

Next, we are interested in the effects of alcohol on a wine's quality rating. Again, the univariate AME of ≈ 0.06 suggests there is a negligible global feature effect. Recall that we motivate FMEs as a local model explanation method first and foremost, which can be extended to regional or global explanations when multiple FMEs are considered. We select a single wine with an ABV of 10.7 (range 8.0 to 14.2) and compute an LLTR for its alcohol content with an NLM threshold value of 0.9. Figure 16 visualizes each explored step size and the corresponding FME and NLM pair. Step sizes that are associated with non-linear effects are greyed out. Indeed, we can observe a large effect on this wine's predicted quality rating given variations in its alcohol content. This confirms our proposition that aggregations of individual FMEs to the AME are not accurately representing feature effects for non-linear models and that evaluating effects for single observations in isolation can provide more insights into the model's workings.

Let us now investigate interactions between both features, first extending our earlier search for an LLTR from Fig. 16 to bivariate step sizes for the same wine, where steps represent 20% of each feature's IQR. We succeed in finding step size combinations

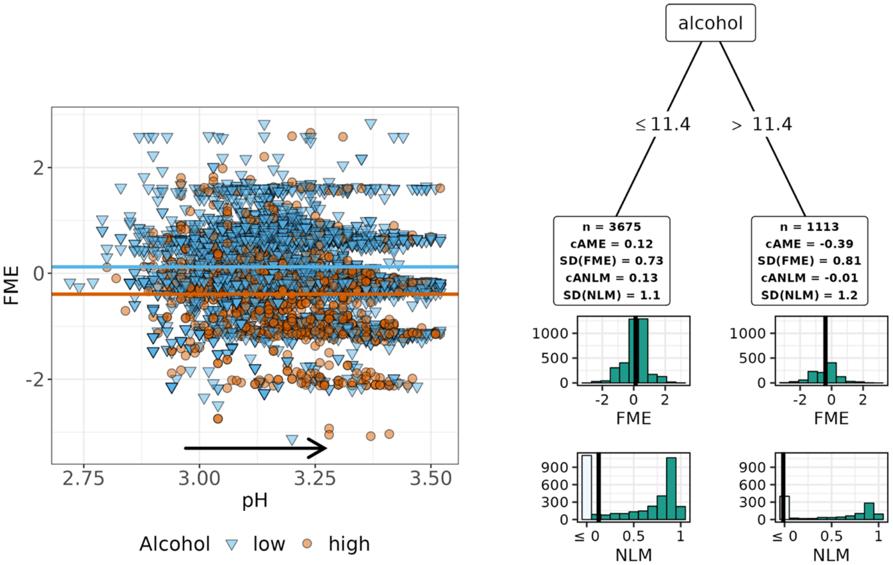


Fig. 15 **White wine data:** FMEs of increasing a wine's pH value by 0.3 on its perceived quality rating, colored by subgroup found by CTREE (left). The colored horizontal lines indicate cAMEs. CTREE finds subgroups whose cAMEs correspond to 0.12 for ABV $\leq$ 11.4 and -0.39 for ABV $>$ 11.4 (right). A similar interaction was found by Goldstein et al. (2015)

that are associated with linear multivariate effects. Next, we evaluate how the data set behaves as a whole, starting with an exploratory analysis of bivariate step sizes and visualizations of FME and NLM distributions via boxplots (see Fig. 18). For combinations of larger step sizes, we can see a large variance in effects. Analyzing the evolution of boxplots through increasing step sizes, we gather that given low pH values, wine quality ratings are driven by the wine's alcohol content (resulting in an increasing dispersion of FMEs for increases in ABV); given high pH values, increasing the alcohol content has a negligible effect on the wine rating (where the dispersion of FMEs for increases in ABV stays roughly the same). Figure 19 visualizes the bivariate distribution of FMEs over both features given a fixed combination of step sizes (+ 0.3 in pH value and +1% in ABV). The largest effects of such a bivariate increase in feature values are mostly located around lower to medium feature value combinations, whereas FMEs are increasingly negative around higher value combinations.

Lastly, we demonstrate how multivariate FMEs can provide insights into the model's workings when other techniques such as ICEs fail, as they are restricted to univariate and bivariate visualizations. In addition to the previous bivariate feature change, we add a 0.5 $\frac{g}{dm^3}$ increase to the potassium sulphate concentration (range 0.22 to 1.08). This noticeably boosts FMEs. In Fig. 20 we visualize the FME density for the threeway feature change and the corresponding AME and ANLM. Again, the AME would obfuscate interpretations by suggesting a negligible effect of this trivariate feature change on the predicted wine quality rating. In contrast to restrictive techniques such as the ICE and PD, we can take advantage of the FME distilling feature effects into a single value for arbitrary feature changes.

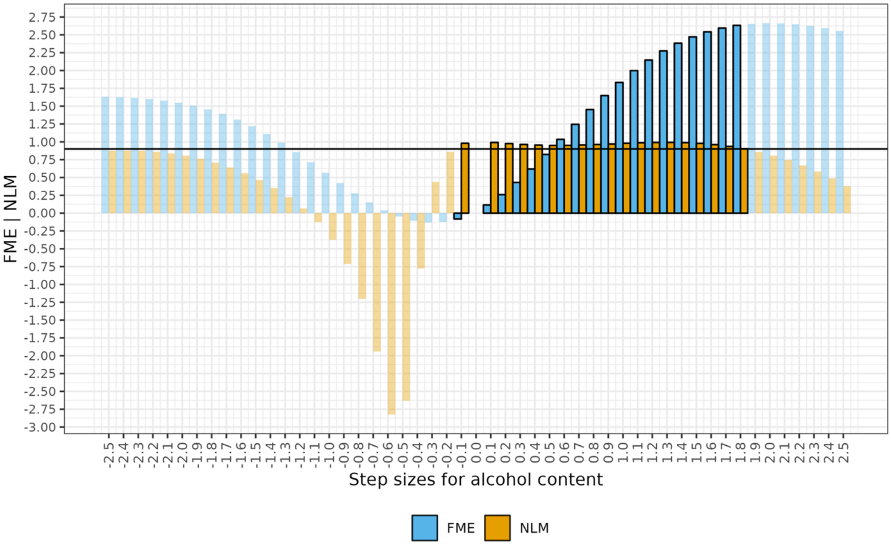


Fig. 16 **White wine data:** Given a single wine (ABV = 10.7), we compute an LLTR (NLM threshold = 0.9) for changes in ABV. Step sizes that are associated with non-linear effects are greyed out

To sum up, we discover that the pH value influences predicted wine quality ratings on a global scale and that the effect differs depending on a wine's alcohol content. ABV has large local effects on predicted wine quality ratings, which cancel each other out when being averaged to an AME. For single observations, we can find trust regions for linear effects. There is an interaction between the pH value and alcohol content with intensely varying effects across observations. The LLTR for ABV can be extended to bivariate changes in pH value and alcohol content for the same, single wine. Furthermore, there are large multimodal effects when adding a third feature change in the potassium sulphate concentration where—again—the AME obfuscates interpretations by indicating a negligible global feature effect.

8 Conclusion

This research paper introduces FMEs as a model-agnostic interpretation method for arbitrary prediction functions, e.g., in the context of ML applications. We create a unified definition of FMEs for both univariate and multivariate, as well as continuous, categorical, and mixed-type features. Furthermore, we introduce an NLM for FMEs based on the similarity between the prediction function and the intersecting linear secant. Due to the complexity and non-linearity of ML models, we suggest to focus on regional instead of global feature effects. We propose a means of estimating expected conditional FMEs via cAMEs and present one strategy to find population subgroups by partitioning the feature space with decision trees. The resulting subgroups can be augmented with cANLM values and CIs in order to receive a compact summary of

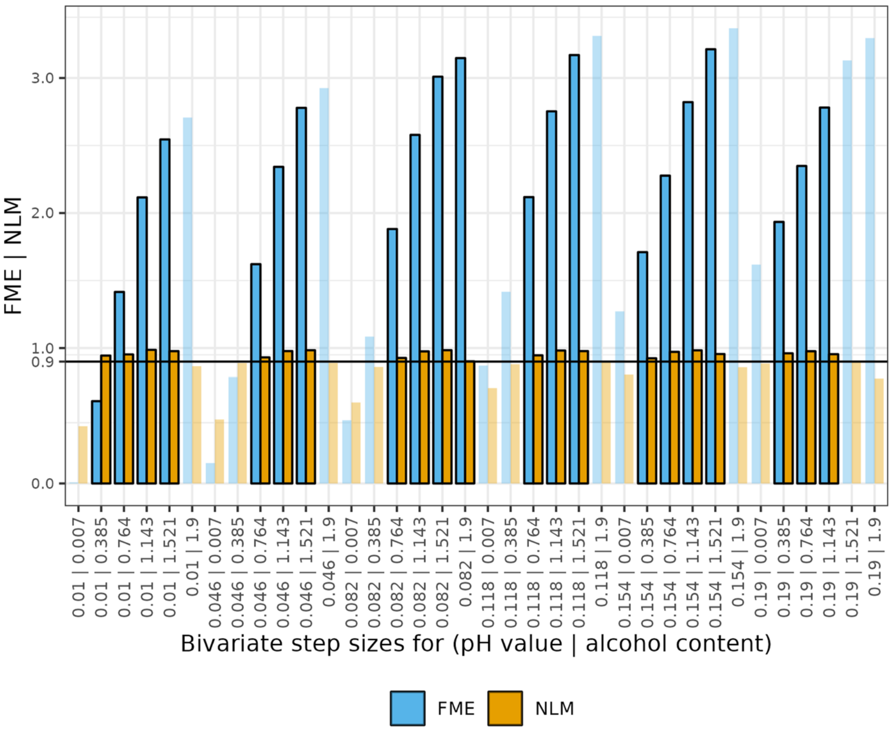


Fig. 17 **White wine data:** We select a single observation (i.e., a single wine) and compute an LLTR for bivariate step size combinations of the pH value and ABV with an NLM threshold of 0.9. Step size combinations that are associated with non-linear effects are greyed out

the prediction function across the feature space. In the Appendix, we provide proofs on the additive recovery property of FMEs and their relation to the ICE and PD.

Given arbitrary predictive models, FMEs can be used to address questions on a model's behavior such as the following: Given pre-specified changes in one or multiple feature values, what is the expected change in predicted outcome? What is the change in prediction for an average observation? What is the change in prediction for a pre-specified observation? What are population subgroups with more homogeneous average effects? What is the degree of non-linearity in these effects? What is our confidence in these estimates? What is the expected change in prediction when switching observed categorical feature values to a reference category?

However, model-agnostic interpretation methods are subject to certain limitations. They are favorable tools to explain the model behavior but often fail to explain the underlying DGP, as the quality of the explanations relies on the closeness between model and reality. Molnar et al. (2022) discuss various general pitfalls of model-agnostic interpretation methods, e.g., model extrapolations, estimation uncertainty, or unjustified causal interpretations.

Throughout the manuscript, we noted various directions that may be explored in future work. For the selection of step sizes, one may work towards better quantifying extrap-

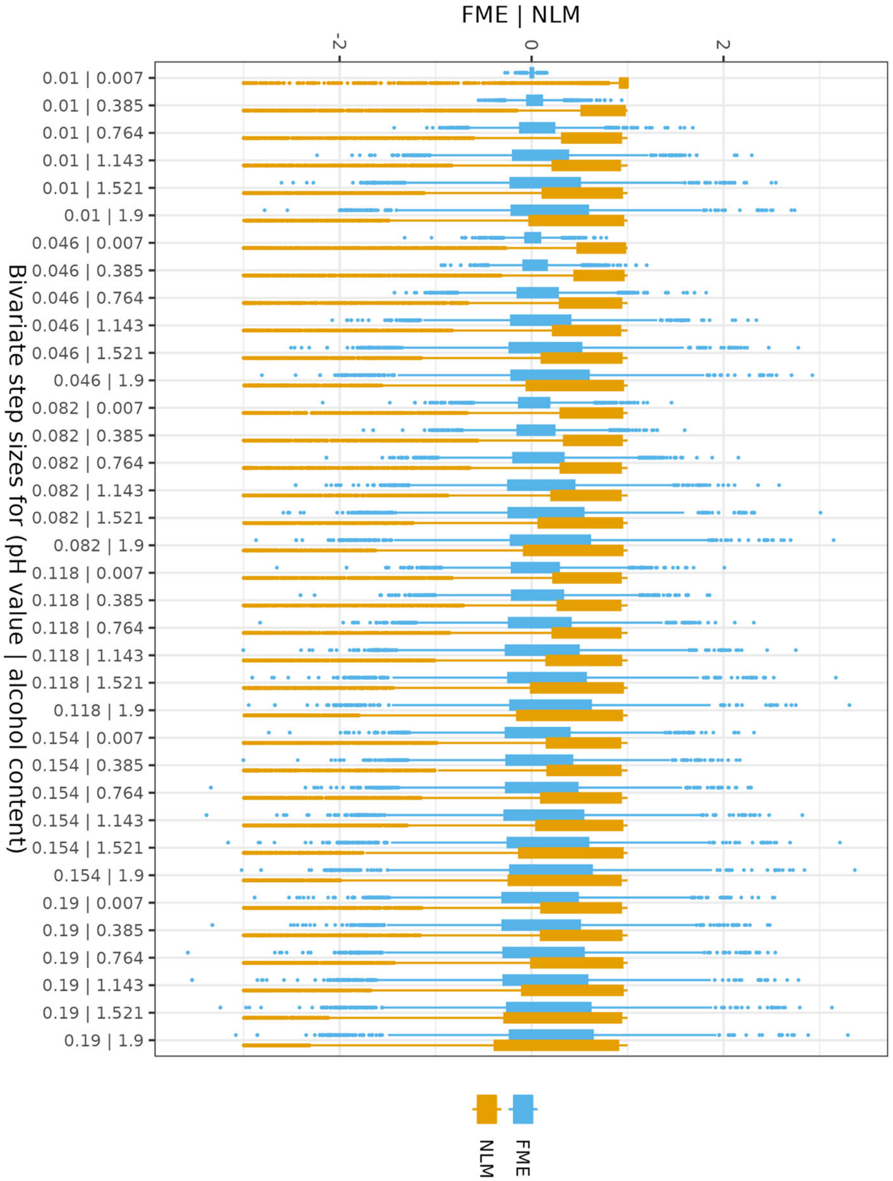


Fig. 18 **White wine data:** Here, we explore how bivariate step sizes affect the global distribution of FME / NLM values. Such an analysis may provide hints about what step sizes or step size combinations drive effects in the model. With visualizations such as in Fig. 19, we can then "zoom in" on a particular step size combination

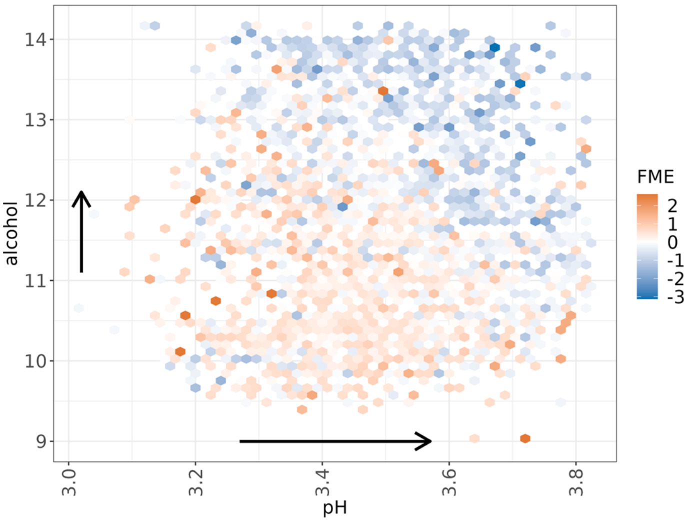


Fig. 19 **White wine data:** Distribution of FMEs given pH and alcohol values. We use averages within hexagons to avoid overplotting values. FME hexagon averages are mostly positive around lower to medium value combinations of both features, while they are increasingly negative around higher value combinations

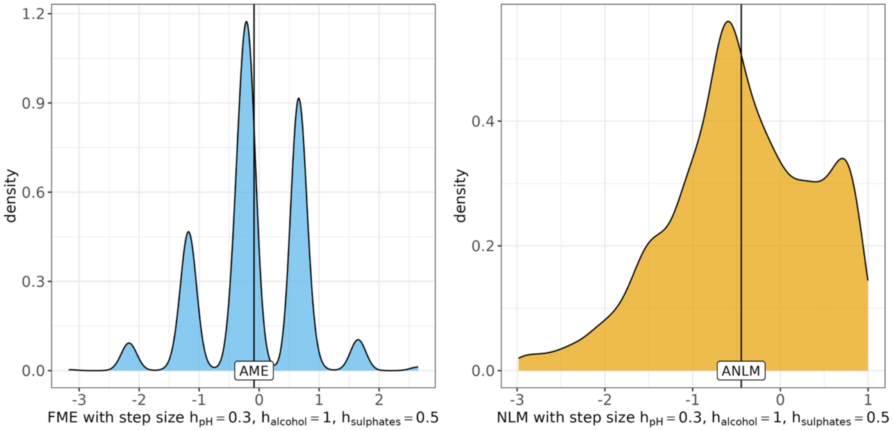


Fig. 20 **White wine data:** Demonstrating how FMEs can be used to interpret the model for threeway interactions when other techniques such as ICEs fail. We evaluate distributions of FMEs for feature changes of the pH value by 0.3, ABV by 1%, and the potassium sulphate concentration by 0.5 $\frac{g}{dm^3}$. FMEs are multimodal. Plotting the corresponding NLM distribution reveals considerable non-linearity for the majority of trivariate FMEs

olation risk. For subgroup selection, one may work towards stabilizing split search or quantifying subgroup uncertainty. To spare computations or facilitate local interpretations, one may search for a subset of representative observations. Furthermore, FMEs may be used for feature importance computations as well.

Many disciplines that have been relying on traditional statistical models—and interpretations in terms of MEs, the AME, MEM, or MER—are starting to utilize the predictive power of ML. With this research paper, we aim to bridge the gap between the restrictive theory on MEs with traditional statistical models and the more flexible and capable approach of interpreting modern ML models with FMEs.

A Background information

A.1 Decomposition of the prediction function

The prediction function to be analyzed may be very complex or even a black box. However, there are multiple ways to decompose the prediction function into a sum of components of increasing order. Although the goal of FMEs is not to decompose the prediction function, it is convenient to either regard the prediction function as an additive decomposition or to keep in mind that it may be decomposed into one. An additive decomposition of the prediction function has the following general form:

$$\widehat{f}(\boldsymbol{x}) = g_{\{0\}} + g_{\{1\}}(x_1) + g_{\{2\}}(x_2) + \cdots + g_{\{1,2\}}(x_1, x_2) + \cdots + g_{\{1,\ldots,p\}}(\boldsymbol{x}) \tag{3}$$

In SA, the additive decomposition is typically referred to as a high-dimensional model representation (HDMR) or ANOVA-HDMR (Saltelli et al. 2008). Various approaches exist to estimate Eq. (3) or a truncated variant, e.g., via recursive computations of PD functions (Hooker 2004b, 2007), random sampling HDMR (Li et al. 2006), or accumulated local effects (Apley and Zhu 2020). Further assumptions are needed to make the decomposition unique, e.g., feature independence (Chastaing et al. 2012). For instance, we may recursively compute Eq. (3) as follows:

$$\begin{aligned}
g_{\{0\}} &= \mathbb{E}_{\boldsymbol{X}}\left[\widehat{f}(\boldsymbol{X})\right] \\
g_{\{1\}}(x_1) &= \mathbb{E}_{\boldsymbol{X}_{-1}}\left[\widehat{f}\,(x_1, \boldsymbol{X}_{-1})\right] - g_{\{0\}} \\
g_{\{2\}}(x_2) &= \mathbb{E}_{\boldsymbol{X}_{-2}}\left[\widehat{f}\,(x_2, \boldsymbol{X}_{-2})\right] - g_{\{0\}} \\
g_{\{1,2\}}(x_1, x_2) &= \mathbb{E}_{\boldsymbol{X}_{-\{1,2\}}}\left[\widehat{f}\left(x_1, x_2, \boldsymbol{X}_{-\{1,2\}}\right)\right] - g_{\{2\}}(x_2) - g_{\{1\}}(x_1) - g_{\{0\}} \\
&\ \ \vdots \\
g_{\{1,\ldots,p\}}(\boldsymbol{x}) &= \widehat{f}(\boldsymbol{x}) - \cdots - g_{\{1,2\}}(x_1, x_2) - g_{\{2\}}(x_2) - g_{\{1\}}(x_1) - g_{\{0\}}
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{X}_{-S}}\left[\widehat{f}(\boldsymbol{x}_S, \boldsymbol{X}_{-S})\right]$ is typically referred to as the PD of $\widehat{f}$ on feature set S in ML. Model decompositions are frequently used in variance-based SA. We refer the reader to the overview by Saltelli et al. (2008) for more details.

A.2 Model extrapolation

King and Zeng (2006) define extrapolation as predicting outside the convex hull of the training data. They demonstrate that the task of determining whether a point is located inside the convex hull can be efficiently solved using linear programming. However, the convex hull may be comprised of many empty areas without training observations, especially in the case of correlated and high-dimensional data. Therefore, it seems plausible to define model extrapolation differently, e.g., as predictions in areas of the feature space with a low density of training points. Hooker (2004a) summarizes two main predicaments of model extrapolations. First, the model creates predictions which do not accurately reflect the target distribution given the features. Second, the predictions are subject to a high variance. Many model-agnostic techniques are subject to model extrapolation risks (Molnar et al. 2022). Hooker (2007) warns against model extrapolations when computing model decompositions. Hooker et al. (2021) call attention to the perils of permuting feature values for feature importance computations. It is important to note that this issue highly depends on the behavior of the chosen model. The issue of determining whether the model extrapolates essentially boils down to quantifying the prediction uncertainty. Some models might diverge considerably from a scenario where they would have been supplied with enough training data (high prediction uncertainty), while other models might be relatively robust against such issues (low prediction uncertainty). Although FMEs based on model extrapolations are still correct in terms of the model output, they might not represent any underlying DGP in an accurate way. Therefore, it is important to take into account (and preferably avoid) potential model extrapolations when selecting feature values and step sizes to compute FMEs.

For some models, built-in measures exist to quantify the prediction uncertainty (Munson and Kegelmeyer 2013), e.g., the proximity measure for tree ensembles which counts how often a pair of points is located in the same leaf node for all trees of the ensemble (Hastie et al. 2001). The same can be done for the pairwise proximity between points in the training and the test set. For instance, given n training observations and a test observation $\boldsymbol{x}$, we can create an $(n \times 1)$ vector of proximities which can be used to detect model extrapolations. However, it is desirable to detect model extrapolations via auxiliary extrapolation risk metrics (AERM) (Munson and Kegelmeyer 2013) which are independent of the trained model. Detecting an EP is similar in concept to the detection of outliers. Although a unified definition of outliers does not exist, they are generally considered to differ as much from other observations as to suspect they were generated by a different mechanism (Hawkins 1980). We can therefore consider an outlier to be drawn from a different distribution than the training data (and one that does not overlap with it), which suits our definition of EPs. In clustering, outliers are often found using local density-based outlier scores such as local outlier probabilities (LOP) (Kriegel et al. 2009). Based on the nearest data points, LOP provides an interpretable score on the scale [0, 1], indicating the probability of a point being an outlier. However, clustering techniques such as LOP are often based on the assumption that the data exhibits a structure of clusters or on assumptions about the clusters' distributions. In theory, one could use various other outlier detection (also referred to

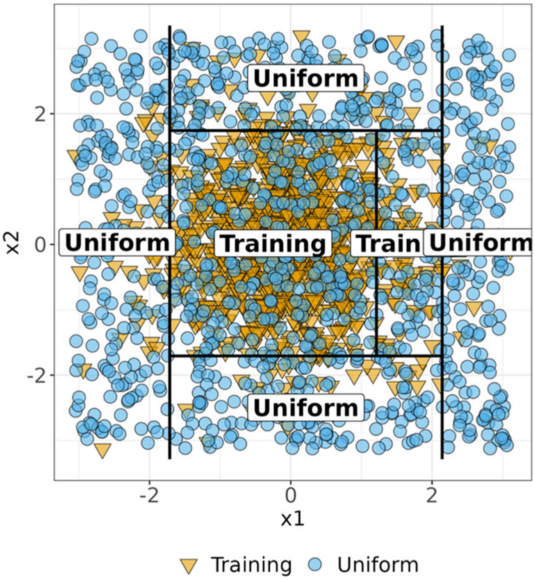


Fig. 21 We augment the training data (orange) with uniform points (blue). A classification tree partitions the feature space into non-extrapolation areas (predominantly occupied with training observations) and extrapolation areas (predominantly occupied with uniform Monte-Carlo samples)

as anomaly detection) mechanisms for extrapolation detection, e.g., isolation forests (Liu et al. 2012).

Hooker (2004a) proposes a statistical test to classify a point as an EP or non-EP. It tests whether a point was more likely to be drawn from the data distribution (non-EP) or the uniform distribution (EP). The uniform distribution is used as an uninformative baseline distribution. The extrapolation risk indicator $R(\boldsymbol{x})$ corresponds to:

$$R(\boldsymbol{x}) = \frac{U(\boldsymbol{x})}{U(\boldsymbol{x}) + P(\boldsymbol{x})} \tag{4}$$

with $U(\boldsymbol{x})$ being the density function of the uniform distribution and $P(\boldsymbol{x})$ the density function of the data distribution. $R(\boldsymbol{x})$ has a range of $[0, 1]$ with 0 indicating the lowest and 1 the highest extrapolation risk. $R(\boldsymbol{x}) > 0.5$ indicates extrapolation. As the support of $U(\boldsymbol{x})$ we may either choose the recommendations of an application domain expert or the observed feature ranges. Equation (4) cannot be directly computed, as the density of the training data is unknown. If $\boldsymbol{x}$ falls outside the multivariate envelope of the training data, it is plausible to set $R(\boldsymbol{x})$ to 1.

We may estimate Eq. (4) by creating a binary classification problem on a data set augmented with uniform Monte-Carlo samples (Hooker 2004a). The training data is labeled as the foreground class. Next, artificial data points are sampled from a uniform distribution and labeled as the background class. A predictive model is trained on the augmented data set and predicts for a given point whether it is more probable that it was drawn from the data distribution or the uniform distribution. Consider two independent standard normally distributed features. We augment the training data with a uniform Monte-Carlo sample with support $[min(x_1), max(x_1)] \times [min(x_2), max(x_2)]$ and use CART to partition the feature space into extrapolation areas and non-extrapolation areas (see Fig. 21). Some training points are located outside the center rectangles in a low-density end of the bivariate normal distribution. Therefore, it is correct to be cautious when evaluating predictions in this area, even if a point was drawn from the training data.

Hooker (2004a) argues that in high-dimensional settings, the Monte-Carlo sample will leave lots of areas of the feature space unoccupied which results in poor classification performance. Classification performance may be boosted by directly utilizing distributional information about the uniform distribution instead of a Monte-Carlo sample. This technique termed confidence and extrapolation representation trees (CERT) exploits a property of classification trees which lets one replace the number of Monte-Carlo points per subspace with the expected number of uniform points at each split. Given the feature space $\mathscr{X}$ with n observations and a subspace $\mathscr{X}_{[\,]}$ with $n_{[\,],\text{data}}$ observations, the expected number of uniform points on the subspace $n_{[\,],\text{ uniform}}$ is proportional to the fraction of feature space hypervolume the subspace occupies:

$$n_{[\,],\text{ uniform}} = \frac{\text{hypervolume}(\mathscr{X}_{[\,]})}{\text{hypervolume}(\mathscr{X})} \cdot n_{[\,],\text{ data}}$$

For the tree growing and pruning strategy, CERT uses a mixture of both CART (e.g., splitting based on the Gini index) and C4.5 (Quinlan 1993) (e.g., missing values and surrogate splits). Apart from letting us directly supply the classification tree with distributional information instead of data, its interpretability is advantageous. The tree partitions the entire feature space at once into hyperrectangles that indicate extrapolation or non-extrapolation areas. Hooker (2004a) argues that CERT provides a markedly lower misclassification rate as opposed to using Monte-Carlo samples with a classification tree. However, it is unclear whether this advantage holds for other classification algorithms used with Monte-Carlo samples.

A.3 Marginal effects for tree-based prediction functions

DMEs are not suited to interpret piecewise constant prediction functions, e.g., classification and regression trees (CART) or tree ensembles such as random forests or gradient boosted trees. Generally, most observations are located on piecewise constant parts of the prediction function where the derivative equals zero. FMEs provide two advantages when interpreting tree-based prediction functions: First, a large enough step size will often involve traversing a jump discontinuity (which corresponds to a tree split in RP) on the prediction function (see Fig. 22), so the FME does not equal zero; second, measures of spread such as the variance can indicate what fraction of FMEs traversed a jump discontinuity and what fraction did not.

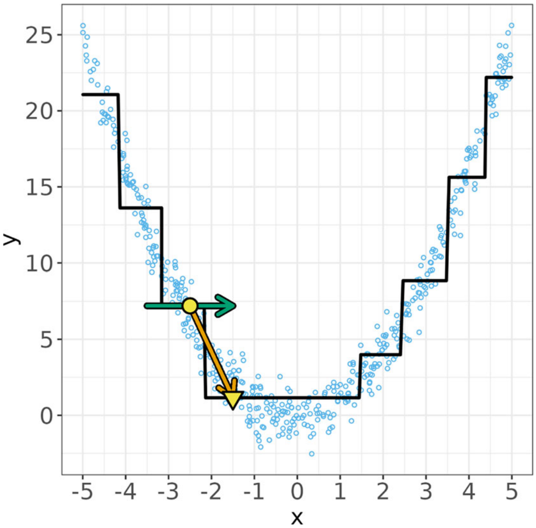


Fig. 22 A quadratic relationship between the target y and a single feature x. A decision tree fits a piecewise constant prediction function (black line) to the training data (blue points). The DME (slope of green arrow) at the point $x = -2.5$ (yellow dot) is zero, while the FME with $h = 1$ traverses the jump discontinuity (secant = orange arrow) and reaches the point $x = -1.5$ (yellow triangle)

B Proofs

B.1 Additive recovery

We provide several proofs on additive recovery based on a prediction function in additive form. Any prediction function can be decomposed into a sum of effect terms of various orders (see Appendix 1). The sum of effect terms of a feature set K is denoted by $\Theta_K(\boldsymbol{x}_K)$. For notational simplicity, the union $\{j\} \cup K$ of the j-th feature index and the index set K is denoted by $\{j, K\}$. The sum of effect terms is denoted by $\Theta_{\{j,K\}}(x_j, \boldsymbol{x}_K)$.

Theorem 1 *(Additive Recovery of Finite Difference) An FD w.r.t. x_j only recovers terms that depend on x_j and no terms that exclusively depend on $\boldsymbol{x}_{-j}$.*

Proof Consider a prediction function $\widehat{f}$ that consists of a sum, including the main effect of x_j, denoted by $g_{\{j\}}(x_j)$, a sum of higher order terms (interactions) between x_j and other features $\boldsymbol{x}_K$, denoted by $\Theta_{\{j,K\}}(x_j, \boldsymbol{x}_K)$, and terms that depend on the remaining features $x_{-\{j,K\}}$, denoted by $\Theta_{-\{j,K\}}(\boldsymbol{x}_{-\{j,K\}})$:

$$\widehat{f}(\boldsymbol{x}) = g_{\{j\}}(x_j) + \Theta_{\{j,K\}}(x_j, \boldsymbol{x}_K) + \Theta_{-\{j,K\}}(\boldsymbol{x}_{-\{j,K\}})$$

It follows that the FD of predictions corresponds to a function that only depends on x_j, i.e., it locally recovers the relevant terms on the interval $[x_j + a, x_j + b]$.

$$\begin{aligned}
FD_{j,\boldsymbol{x},a,b} &= \widehat{f}(x_1, \ldots, x_j + a, \ldots, x_p) - \widehat{f}(x_1, \ldots, x_j + b, \ldots, x_p) \\
&= \left[g_{\{j\}}(x_j + a) \ + \ \Theta_{\{j,K\}}(x_j + a, \boldsymbol{x}_K) \ + \ \Theta_{-\{j,K\}}(\boldsymbol{x}_{-\{j,K\}})\right] \\
&\quad - \left[g_{\{j\}}(x_j + b) \ + \ \Theta_{\{j,K\}}(x_j + b, \boldsymbol{x}_K) \ + \ \Theta_{-\{j,K\}}(\boldsymbol{x}_{-\{j,K\}})\right] \\
&= g_{\{j\}}(x_j + a) - g_{\{j\}}(x_j + b) + \Theta_{\{j,K\}}(x_j + a, \boldsymbol{x}_K) \\
&\quad - \Theta_{\{j,K\}}(x_j + b, \boldsymbol{x}_K)
\end{aligned}$$

□

Corollary 1 *(Additive Recovery of Univariate Forward Marginal Effect) The univariate FME w.r.t. x_j only recovers terms that depend on x_j and no terms that exclusively depend on $\boldsymbol{x}_{-j}$.*

Proof Consider a prediction function $\widehat{f}$ that consists of a sum, including the main effect of x_j, denoted by $g_{\{j\}}(x_j)$, a sum of higher order terms (interactions) between x_j and other features $\boldsymbol{x}_K$, denoted by $\Theta_{\{j,K\}}(x_j, \boldsymbol{x}_K)$, and terms that depend on the remaining features $\boldsymbol{x}_{-\{j,K\}}$, denoted by $\Theta_{-\{j,K\}}(\boldsymbol{x}_{-\{j,K\}})$:

$$\widehat{f}(\boldsymbol{x}) = g_{\{j\}}\left(x_j\right) + \Theta_{\{j,K\}}\left(x_j, \boldsymbol{x}_K\right) + \Theta_{-\{j,K\}}\left(\boldsymbol{x}_{-\{j,K\}}\right)$$

The FD w.r.t. x_j is equivalent to the FME w.r.t. x_j with $a = h_j$ and $b = 0$. Using Theorem 1, it follows that:

$$\text{FME}_{\boldsymbol{x},h_j} = g_{\{j\}}\left(x_j + h_j\right) - g_{\{j\}}\left(x_j\right) + \Theta_{\{j,\,K\}}\left(x_j + h_j, \boldsymbol{x}_K\right) - \Theta_{\{j,\,K\}}\left(x_j, \boldsymbol{x}_K\right)$$

□

Theorem 2 *(Additive Recovery of Multivariate Forward Marginal Effect) The multivariate FME w.r.t. $\boldsymbol{x}_S$ only recovers terms that depend on $\boldsymbol{x}_S$ and no terms that exclusively depend on $\boldsymbol{x}_{-S}$.*

Proof Consider a feature set S. The power set of S excluding the empty set is denoted by $\mathscr{P}^* = \mathscr{P}(S) \setminus \emptyset$. The prediction function $\widehat{f}$ consists of a sum, including the sum of effects of all subsets of features $K \in \mathscr{P}^*$, denoted by $\sum_{K \in \mathscr{P}^*} g_K\left(\boldsymbol{x}_K\right)$, and a sum of terms that depend on the remaining features, denoted by $\Theta_{-S}(\boldsymbol{x}_{-S})$:

$$\begin{aligned}\widehat{f}(\boldsymbol{x}) &= \sum_{K \in \mathscr{P}^*} g_K\left(\boldsymbol{x}_K\right) + \Theta_{-S}\left(\boldsymbol{x}_{-S}\right)\\ \mathrm{FME}_{\boldsymbol{x},\boldsymbol{h}_S} &= \left[\sum_{K \in \mathscr{P}^*} g_K(\boldsymbol{x}_K + \boldsymbol{h}_K) + \Theta_{-S}(\boldsymbol{x}_{-S})\right]\\ &\quad - \left[\sum_{K \in \mathscr{P}^*} g_K(\boldsymbol{x}_K) + \Theta_{-S}\left(\boldsymbol{x}_{-S}\right)\right]\\ &= \sum_{K \in \mathscr{P}^*} \left[g_K(\boldsymbol{x}_K + \boldsymbol{h}_K) - g_K(\boldsymbol{x}_K)\right]\end{aligned}$$

□

B.2 Relation between forward marginal effects, the individual conditional expectation, and partial dependence

Theorem 3 *(Equivalence between Forward Marginal Effect and Forward Difference of Individual Conditional Expectation) The FME with step size $\boldsymbol{h}_S$ is equivalent to the forward difference with step size $\boldsymbol{h}_S$ between two locations on the ICE.*

Proof

$$\begin{aligned}\mathrm{FME}_{\boldsymbol{x},\boldsymbol{h}_S} &= \widehat{f}(\boldsymbol{x}_S + \boldsymbol{h}_S, \boldsymbol{x}_{-S}) - \widehat{f}(\boldsymbol{x})\\ &= \mathrm{ICE}_{\boldsymbol{x},S}(\boldsymbol{x}_S + \boldsymbol{h}_S) - \mathrm{ICE}_{\boldsymbol{x},S}(\boldsymbol{x}_S)\end{aligned}$$

□

Theorem 4 *(Equivalence between Average Marginal Effect and Forward Difference of Partial Dependence for Linear Prediction Functions) The AME with step size $\boldsymbol{h}_S$ is equivalent to the forward difference with step size $\boldsymbol{h}_S$ between two locations on the PD for prediction functions that are linear in $\boldsymbol{x}_S$.*

Proof If $\widehat{f}$ is linear in $\boldsymbol{x}_S$:

$$\begin{aligned}&\widehat{f}\left(\boldsymbol{x}_S{}^{(i)} + \boldsymbol{h}_S\right) = \widehat{f}(\boldsymbol{x}_S + \boldsymbol{h}_S) \quad \forall \ i \ \in \ \{1, \ldots, n\},\\ &\boldsymbol{x}_S, \ \boldsymbol{h}_S \ \in \ \times_{j \in S} \mathscr{X}_j\end{aligned} \tag{5}$$

It follows:

$$\begin{aligned}
\text{AME}_{\mathscr{D},\mathbf{h}_S} &= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{f}\left({\mathbf{x}_S}^{(i)}+\mathbf{h}_S,{\mathbf{x}_{-S}}^{(i)}\right)-\widehat{f}\left(\mathbf{x}^{(i)}\right)\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\widehat{f}\left({\mathbf{x}_S}^{(i)}+\mathbf{h}_S,{\mathbf{x}_{-S}}^{(i)}\right)-\frac{1}{n}\sum_{i=1}^{n}\widehat{f}\left({\mathbf{x}_S}^{(i)},{\mathbf{x}_{-S}}^{(i)}\right)\\
&\overset{(5)}{=}\frac{1}{n}\sum_{i=1}^{n}\widehat{f}\left(\mathbf{x}_S+\mathbf{h}_S,{\mathbf{x}_{-S}}^{(i)}\right)-\frac{1}{n}\sum_{i=1}^{n}\widehat{f}\left(\mathbf{x}_S,{\mathbf{x}_{-S}}^{(i)}\right)\\
&= \widehat{\text{PD}}_{\mathscr{D},S}\left(\mathbf{x}_S+\mathbf{h}_S\right)-\widehat{\text{PD}}_{\mathscr{D},S}\left(\mathbf{x}_S\right)
\end{aligned}$$

□

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability All data are created or provided in the following public repository: https://github.com/scholbeck/forward_marginal_effects.git

Code availability We provide reproducible scripts for our simulations and the applied example in the following public repository: https://github.com/scholbeck/forward_marginal_effects.git

Declarations

Conflict of interest Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable.

References

Ai C, Norton EC (2003) Interaction terms in logit and probit models. Economics Letters 80(1):123–129

Alt H, Godau M (1995) Computing the Fréchet distance between two polygonal curves. International Journal of Computational Geometry & Applications 05(01n02):75–91

Ancona M, Ceolini E, Öztireli C, Gross M (2018) Towards better understanding of gradient-based attribution methods for deep neural networks. In: International Conference on Learning Representations, https://openreview.net/forum?id=Sy21R9JAW

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(4):1059–1086

Arel-Bundock V (2023) marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests. https://marginaleffects.com/, R package version 0.15.1.9002

Athey S (2017) Beyond prediction: Using big data for policy problems. Science 355(6324):483–485

Bartus T (2005) Estimation of marginal effects using margeff. The Stata Journal 5(3):309–329

Belogay E, Cabrelli C, Molter U, Shonkwiler R (1997) Calculating the Hausdorff distance between curves. Information Processing Letters 64(1):17–22

Bertsimas D, Dunn J (2017) Optimal classification trees. Machine Learning 106(7):1039–1082

Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Breiman L (2001b) Statistical modeling: The two cultures. Statist Sci 16(3):199–231, with comments and a rejoinder by the author

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA

Casalicchio G, Molnar C, Bischl B (2019) Visualizing the feature importance for black box models. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science, Springer, Cham, vol 11051

Chastaing G, Gamboa F, Prieur C (2012) Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. Electronic Journal of Statistics 6:2420–2448

Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009) Wine Quality. UCI Machine Learning Repository, https://doi.org/10.24432/C56S3T

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20(177):1–81

Friedman JH (1991) Multivariate Adaptive Regression Splines. The Annals of Statistics 19(1):1–67

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5):1189–1232

Gelman A, Pardoe I (2007) Average predictive comparisons for models with nonlinearity, interactions, and variance components. Sociological Methodology 37(1):23–51

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24(1):44–65

Greene W (2012) Econometric Analysis. Pearson International Edition, Pearson Education Limited

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc

Hawkins DM (1980) Identification of Outliers. Springer, Netherlands,. https://doi.org/10.1007/978-94-015-3994-4_1

Hooker G (2004a) Diagnosing extrapolation: Tree-based density estimation. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '04, p 569-574

Hooker G (2004b) Discovering additive structure in black box functions. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '04, pp 575–580

Hooker G (2007) Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. Journal of Computational and Graphical Statistics 16(3):709–732

Hooker G, Mentch L, Zhou S (2021) Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. Statistics and Computing 31(6):82

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3):651–674

King G, Zeng L (2006) The dangers of extreme counterfactuals. Political Analysis 14(2):131–159

Kriegel HP, Kröger P, Schubert E, Zimek A (2009) LoOP: Local outlier probabilities. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, CIKM '09, p 1649-1652

Last M, Maimon O, Minkov E (2002) Improving stability of decision trees. International Journal of Pattern Recognition and Artificial Intelligence 16(02):145–159

Leeper TJ (2018) margins: Marginal effects for model objects. https://CRAN.R-project.org/package=margins, R package version 0.3.23

Li G, Hu J, Wang SW, Georgopoulos PG, Schoendorf J, Rabitz H (2006) Random sampling-high dimensional model representation (RS-HDMR) and orthogonality of its different order component functions. The Journal of Physical Chemistry A 110(7):2474–2485

Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. ACM Trans Knowl Discov Data 6(1)

Loh WY (2014) Fifty years of classification and regression trees. International Statistical Review 82(3):329–348

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp 4765–4774

Löwe H, Scholbeck CA, Heumann C, Bischl B, Casalicchio G (2023) fmeffects: An R package for forward marginal effects. arXiv e-prints arXiv:2310.02008

Mize TD, Doan L, Long JS (2019) A general framework for comparing predictions and marginal effects across models. Sociological Methodology 49(1):152–189

Molnar C (2022) Interpretable Machine Learning, 2nd edn. https://christophm.github.io/interpretable-ml-book

Molnar C, Casalicchio G, Bischl B (2020) Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: Cellier P, Driessens K (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, vol 1167, Springer, Cham

Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, Casalicchio G, Grosse-Wentrup M, Bischl B (2022) General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200, Springer, Cham

Morris MD (1991) Factorial sampling plans for preliminary computational experiments. Technometrics 33(2):161–174

Mullahy J (2017) Marginal effects in multivariate probit models. Empirical economics 53(2):447–461

Munson MA, Kegelmeyer WP (2013) Builtin vs. auxiliary detection of extrapolation risk. Tech. rep., Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California

Norouzi M, Collins MD, Johnson M, Fleet DJ, Kohli P (2015) Efficient non-greedy optimization of decision trees. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press, Cambridge, MA, USA, NIPS'15, p 1729-1737

Norton EC, Dowd BE, Maciejewski ML (2019) Marginal effects-quantifying the effect of changes in risk factors in logistic regression models. JAMA 321(13):1304–1305

Onukwugha E, Bergtold J, Jain R (2015) A primer on marginal effects-part II: Health services research applications. PharmacoEconomics 33(2):97–103

Philipp M, Zeileis A, Strobl C (2016) A toolkit for stability assessment of tree-based learners. In: Proceedings of COMPSTAT 2016 - 22nd International Conference on Computational Statistics, The International Statistical Institute/International Association for Statistical Computing, p 315-325

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California

Ramsey SM, Bergtold JS (2021) Examining inferences from neural network estimators of binary choice processes: Marginal effects, and willingness-to-pay. Computational Economics 58(4):1137–1165

Razavi S, Gupta HV (2016) A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resources Research 52(1):423–439

Razavi S, Jakeman A, Saltelli A, Prieur C, Iooss B, Borgonovo E, Plischke E, Lo Piano S, Iwanaga T, Becker W, Tarantola S, Guillaume JH, Jakeman J, Gupta H, Melillo N, Rabitti G, Chabridon V, Duan Q, Sun X, Smith S, Sheikholeslami R, Hosseini N, Asadzadeh M, Puy A, Kucherenko S, Maier HR (2021) The future of sensitivity analysis: An essential discipline for systems modeling and policy support. Environmental Modelling and Software 137:104954

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '16, p 1135-1144

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global Sensitivity Analysis: The Primer. John Wiley & Sons, Ltd

Scholbeck CA, Molnar C, Heumann C, Bischl B, Casalicchio G (2020) Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. In: Cellier P, Driessens K (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, vol 1167, Springer, Cham

Seibold H, Zeileis A, Hothorn T (2016) Model-based recursive partitioning for subgroup analyses. The International Journal of Biostatistics 12(1):45–63

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '20, p 180-186

Sobol I, Kucherenko S (2010) Derivative based global sensitivity measures. Procedia - Social and Behavioral Sciences 2(6):7745 – 7746, Sixth International Conference on Sensitivity Analysis of Model Output

Stachl C, Hilbert S, Au JQ, Buschek D, De Luca A, Bischl B, Hussmann H, Bühner M (2017) Personality traits predict smartphone usage. European Journal of Personality 31(6):701–722

StataCorp, (2023) Stata Statistical Software: Release 18. StataCorp LLC, College Station, TX

Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41(3):647–665

Turney P (1995) Technical note: Bias and the quantification of stability. Machine Learning 20(1):23–33

Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology 31(2):841–887

Williams R (2012) Using the margins command to estimate and interpret adjusted predictions and marginal effects. Stata Journal (24) 12(2):308–331

Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17(2):492–514

Zhao X, Yan X, Yu A, Van Hentenryck P (2020) Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel Behaviour and Society 20:22–35

Zhou Y, Zhou Z, Hooker G (2023) Approximation trees: Statistical reproducibility in model distillation. Data Mining and Knowledge Discovery. https://doi.org/10.1007/s10618-022-00907-3

Zhou Z, Hooker G, Wang F (2021) S-LIME: Stabilized-lime for model explanation. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2429-2438